

Friday, October 31. 2008

## **Solaris 10 10/08 - out now**

The download page has changed, so you can download it now at the usual locations at the sun website

Posted by Joerg Moellenkamp in Solaris at 14:51

## **What's new in Solaris 10 10/08 (also known as Update 6)**

The "What's new in What's New in the Solaris 10 10/08 Release" document is online at docs.sun.com. Thus you can get an overview what you can expect from the next release of Solaris 10.

I will provide a list of the features for you in this article. For more informations please consult the linked article. As you see, there was a lot of development in the ZFS part of Solaris 10. Another really interesting feature is the update-on-attach for Zones. This solves some problems like the migration of zones from normal SPARC to a CMT SPARC system (sun4u versus sun4v).

Okay, in Solaris 10/08 aka Update 6 you will find:

### Installation Enhancements

- Solaris Installation for ZFS Root Pools

### System Administration Enhancements

#### ZFS Command Improvements and Changes

- ZFS installation and boot support
- Rolling back a ZFS dataset without unmounting
- Enhancements to the zfs send command
- ZFS quotas and reservations for file system data only
- new ZFS storage pool properties
- ZFS command history enhancements
- support for upgrading ZFS filesystems
- ZFS delegated administration
- Setting up separate ZFS logging devices
- Creating intermediate ZFS datasets
- ZFS hot-plugging enhancements
- GZIP compression now available for ZFS
- Storing multiple copies of ZFS user data

#### Solaris Installation Tool Support of ZFS File Systems

- Solaris interactive text installer to install a UFS or a ZFS root file system.

- Custom JumpStart features to set up a profile to create a ZFS storage pool and designate a bootable ZFS file system.

- Migrate a UFS root file system to a ZFS root file system by using the Solaris Live Upgrade feature.

- Set up a mirrored ZFS root pool by selecting two disks during the installation.

- Automatically create swap and dump devices on ZFS volumes in the ZFS root pool.

#### SunVTS 7.0 Patch Set 3

- lockstat Provider for DTrace. DTrace lockstat probes that displayed the spin count (spins) now returns spin time in nanoseconds.

### System Resource Enhancements

#### New Solaris Zones Features

- Update on Attach.If the new host has the same or later versions of the zone-dependent packages and their

associated patches, using zoneadm attach with the -u option, updates those packages within the zone to match the new host.[...] This option also enables automatic migration between machine classes, such as from sun4u to sun4v.

Ability to Set Default Router in Shared-IP Zone

ZFS Zone Path Permitted

x86: New GRUB findroot Command

x64: Support for 256 Processors

#### System Performance Enhancements

SPARC: Solaris SPARC Boot Architecture Redesigned

x86: Kernel Support for Intel SSSE3, SSE4.1, SSE4.2, and AMD SSE4A

#### Security Enhancement

Separation of Duty Enforcement Through the Solaris Management Console

SHA256/SHA512 crypt(3C) Plug-in

pam\_list Module

#### Desktop Enhancements

SPARC: Adobe Reader 8.1.2

Flash Player 9.0.124.0

#### Networking Enhancements

Communication Protocol Parser Utilities

SIP End-to-end Traffic Measurements and Logging

#### Device Management Enhancements

Faulty Device Retirement Feature

MPxIO Support for Hitachi Adaptable Modular Storage Series Arrays

#### Driver Enhancements

x86: NVIDIA ck804/mcp55 SATA Controller Driver

x86: LSI MegaRAID SAS Controllers Driver

ixgbe Driver. The ixgbe is a 10 Gigabit PCI Express Ethernet driver that supports Intel 82598 10 Gigabit Ethernet controller.

SPARC: Support for aac Driver

#### Additional Software Enhancements

Perl Database Interface and Perl PostgreSQL Driver

PostgreSQL 8.3

#### Language Support Enhancements

IIMF Hangul Language Engine. The Hangul LE (Language Engine) is a new Korean input method.

#### Freeware Enhancements

C-URL - The C-URL Wrappers Library

Libidn - Internationalized Domain Library

LibGD - The Graphics Draw Library

TIDY HTML Library

You will find Solaris 10/08 for download at the usual locations, but at the moment there is still the 05/08 release.

Posted by Joerg Moellenkamp in Solaris at 13:14

**links for 2008-10-31**

Oracle parallel query performance on a T5140 - The herbal guide to holistic troubleshooting

(tags: solaris oracle)

the storage anarchist: 1.028: benchmarking. badly.

About benchmarking in storage. Albeit heavily biased to EMC, it's a good read how IBM and TMS reached their 1.000.000 IOPS records

(tags: storage benchmarking)

Dimitri (dim) Tools HOMEPAGE

(tags: tools sysadmin solaris monitoring performance)

dimSTAT by examples - Roman Ivanov

(tags: analysis tuning)

InfoQ: JavaOne: Garbage First

(tags: jvm java garbagecollection)

Posted by del.icio.us at 12:00

Thursday, October 30, 2008

### **Jonathan about Q1FY09**

Jonathan wrote an interesting article about his perspective on the first quarter of the current fiscal year: Understanding Sun's Business - Q1 Results. Really worth a read.

Posted by Joerg Moellenkamp in Sun at 22:19

### **Ouch ....**

On our first quarter of this fiscal year: Net loss for the first quarter of fiscal 2009 on a GAAP basis was \$1.677 billion, or \$(2.24) per share on a diluted basis, as compared with a net income of \$89 million, or \$0.10 per share, for the first quarter of fiscal 2008. GAAP net loss per share includes a \$1.445 billion non-cash charge for goodwill impairment. It also includes a restructuring charge of approximately \$63 million pursuant to the restructuring that commenced in the fourth quarter of fiscal 2008.

On a non-GAAP basis, net loss for the first quarter of fiscal 2009 was \$65 million, or \$(0.09) per share on a diluted basis, as compared with a non-GAAP net income of \$285 million, or \$0.32 per share, for the first quarter of fiscal 2008.

Posted by Joerg Moellenkamp in Sun at 21:17

### **links for 2008-10-30**

Explore Your ZFS Adaptic Replacement Cache (ARC) - The Blog of Ben Rockwood

(tags: zfs solaris cache)

Posted by del.icio.us in del.icio.us at 19:00

Wednesday, October 29, 2008

### links for 2008-10-29

Wider der Verschraddelung der IT | Einstiegsserver mit SPARC64-Quadcore-Pro... | iX-Newsforen  
Sorry ... konnte mich mal wieder nicht zurueckhalten ... ich habe mal wieder laenger kommentiert.  
(tags: heise)

Calamity Coach or Thirteen reasons not to travel

(tags: travel inspiration fun humour illustration art comic)

Bush-A\_P\_Address\_Hack\_-\_The\_Revenge\_of\_the\_Stupid\_Core.tKhf.pdf (application/pdf-Objekt)

(tags: isp)

dropsafe : Thunderbird: Just because it's open source doesn't mean it's not shit

(tags: oss bugs)

Posted by del.icio.us in del.icio.us at 12:00

### Memory prices for SPARC Servers.

One of the critics about the M3000 isn't the system itself. It's the price for the memory. But as always, this isn't about making Sun rich. There are technical reasons for this: Let's assume you buy an x86 system. Most of the systems are phased out within 3 years or so. Now think about Suns: I know several customers, that still uses E250 or E450 for certain tasks. Without any problems. This systems were current system 8 years ago. But they still work. The reason for this: Extreme quality standards for components.

It's important to know, that electronics scatter vastly in the fulfillment of their specifications. This is the reason, why there is a frequency number on your processors, on your memory. Because the fulfillment of specification may vary with the frequency or with the temperature or the age. So you test your electronics and print the most expensive frequency in accordance to your specifications on the chip casing (okay, there are some problems with a matured manufacturing, sometimes you don't produce enough low specification modules, so you have to downgrade better parts). This scatter in quality lies in the nature of mass producing electronic parts.

When Sun wants to sell memory, the standards are really rigid. At first, we don't take the memory from the spot market and sell it. At first Sun defines exact specs for the memory. If the memory module doesn't fulfill this specs within a really thin margin ... back to the drawing table for the manufacturer. When the manufacturer fulfill this specs, Sun takes modules from the last few months from all fabs of the manufacturer and check the modules by aging them artificially: A month or two at a temperature above the specs running at the threshold frequency. In a month you can simulate a several years lifetime of the memory module.

When more than really small number of modules fail in this time ... well ... back to the drawing board for the manufacturer. Just when all tests are fulfilled within the really narrow test specification, the memory modules are used for the our SPARC servers.

So: Why do you have to spend so much money for the memory. Well ... if you want DIMMS within +/- 10 percent of the specification you can choose from a vast amount of modules. If you want memory modules within 0.5% of your specification, the choice gets really thin. At 0.1% they are hand selected (this are not the current numers ... numbers just chosen to give you an impression). The same for long time stability: If you want a series of memory modules with a lifetime of 3 years, you get choose from a vast amount of memory modules. At 10 years the story looks really different. So we want the cream of the crop of memory modules in at least two dimensions. And when you want to have the best quality, you have to pay for it. So we can't simply take this el-cheapo DIMMS from the market and put it into a SPARC server. We have to buy the expensive modules from the manufactures, as the manufacturer know as well, that they sell

their top quality to us.

To answer the overarching question: Does high-quality memory modules really matter? Yes ... definitely. Perhaps not for your PC at home. But surely for systems running your business for the next years.

Posted by Joerg Moellenkamp in Sun at 10:42

Tuesday, October 28. 2008

### **Clifford Stoll at TED**

Clifford Stoll was the guy, who found some hackers hired by the KGB in his system 20 years ago. In computer security he is known for his book "The Cuckoo's Egg" (albeit it's more a popular science book), which describes the hunt after the attackers. 2 years ago he held a presentation at TED. His style is a combination of Carl Sagan and Roger Rabbit (as a commentator at the TED site wrote) but it gives you many interesting insights. Really worth a look.

Posted by Joerg Moellenkamp in Fundsache at 17:56

### **Sun SPARC Enterprise M3000 announced**

Sun and Fujitsu jointly announced the system codenamed Ikkaku today : Sun SPARC Enterprise M3000.

The M3000 is a single-socket system for the SPARC 64 VII CPU, thus it's a 4 core system, executing up to 8 threads. Up to 32 GB memory. 4 PCI-Slots. So it's an entry-level system. Starting at \$14.795,00.

But before you compare it with your favourite brand of el-cheapo x86 systems. This system shares many RAS features of his bigger brothers: Instruction Retry on processor level ... ECC for memory ... ECC for the interconnect between system controller and the CPU ... ECC for integer registers ... parity for floatingpoint registers ... fault isolation ... a single defective core doesn't keep the system from starting up, etc. ... well ... just read the architecture whitepaper on your own.

Posted by Joerg Moellenkamp in Sun at 14:08

### **links for 2008-10-28**

Penguin Pete's Blog - Why I Am Not A "Linux Advocate"

(tags: linux advocacy)

Posted by del.icio.us in del.icio.us at 12:00

Monday, October 27. 2008

### **Pedantic nation**

I had my trip back to Hamburg from Toulouse today. Was a nice flight. Well, except the contact with the ground crew in Toulouse. I had the impertinence to try to board while the boarding of rows 12 to 25 took place. The look from the lady was able to kill mere mortals. She didn't allowed me to board the plane. Okay ... waiting 30 seconds. She announced the immediate boarding of all rows. Now i was allowed to enter the aircraft. Dammed! French people are really pedantic for a nation with a revolution in its history ...

Posted by Joerg Moellenkamp in Aviation at 20:55

### **CRM bei der Bahn**

Also irgendwie muss die Bahn noch ein wenig an ihrem CRM System arbeiten. Ich habe meine Nachfolge-Bahncard im Abo erhalten: [...] natuerlich mit allen exklusiven Vorteilen, von denen nur unsere besten Kunden profitieren. Denn ab sofort geniessen Sie den bahn.comfort Status"Hmm ... find ich ja auch toll das ich in die Lounge darf ... aber ab sofort ... ich glaube ich bin schon seit 2003 in diesem Frequently-Delayed-Programm ....

Posted by Joerg Moellenkamp in General at 20:35

### **SAP Benchmarks revisited**

There is a new benchmark from Sun for SAP on the SAP website. SAP certified the benchmarking of the Sun Fire X4600 with quadcores. This benchmark result is especially interesting when you compare it to another 8 socket system - the HP ProLiant DL785. Both systems have the same basic characteristics:

The Sun hardware:Sun Fire X4600M2, 8 processors / 32 cores / 32 threads, Quad-Core AMD Opteron Processor 8360 SE, 2.5 GHz, 128 KB L1 cache and 512 KB L2 cache per core, 2 MB L3 cache per processor, 128 GB main memory  
The HP hardware:HP ProLiant DL785, 8 processors / 32 cores / 32 threads, Quad-Core AMD Opteron processor 8360 SE, 2.5 GHz, 128 KB L1 cache and 512 KB L2 cache per core, 2 MB L3 cache per processor, 128 GB main memory  
The Sun system uses Solaris 10 and MaxDB 7.6, the HP system uses Windows Server 2003 Enterprise Edition and SQL Server 2008.

The Sun system yields 5,800 SD Users and 29,670 SAPS (SAP Benchmark Certificate: 2008061). The HP system yields 5,230 SD Users and 26,180 SAPS (SAP Benchmark Certificate: 2008026). That's quite a difference, but not the end of the story: The Sun result is a 6.0 unicode result, the HP is a plain 6.0 result. You may remember: You loose at least 15% (depends on the systems) performance by using unicode. That's a really impressive benchmark result. It's the fastest 8 socket x86 SAP benchmark result at the moment.

As both system uses different operating systems and different databases, you can't say if it's a bad idea to use SAP on HP or a bad idea to use SAP on Windows/SQL Server

Posted by Joerg Moellenkamp in Sun at 19:51

### **Backscatter**

There is really some buzz about xray backscatter scanners here in Germany. Many newspaper titled with "European Commision allows nude scanners". I reported about them in an article back in 2005 when Bruce "He knows Alice and Bobs shared secret" Schneier wrote about them. At that time i thought about them as a maximum invasive security measure. But i think different now: At first .. the software of this scanners has evolved .. they blur the forms of the body or they entirely deletes it from the image leaving just the foreign objects in the image.

But there is an important additional reason: Let's assume you fly in vacation: Most of the womens take an bathing suit with them, that discloses more than it covers. And the bathing trunks for men aren't especially covering pieces of clothing (even besides the Borat-type ones). There are some functional needs that led to this design of bathing wear. Just thinks about hydrodynamics. It's not a problem to go to beach in clothings like this ... why it's a problem to be

scanned with less amount of details.

Okay. The correct question is: What can i get back for disclosing my body in this manner? That's simple. Today a coin in your pocket, the underwire of a bra or a metal shoelace eyelet can send you a body search. I found it more invasive to be touched and searched by strangers than showing them something they could see on television, in newspapers or at the beach in a much better quality (image quality and quality of the displayed body ) (BTW: There is and reaaaally old joke. Goes like "I like flying ... without bodychecks at the airport i wouldn't have any sexual life"). When we can get rid of body checks in lieu of walking through a backscatter scanner, i'm okay with that.

It's pretty much an hyped discussion at the moment. Hope there will be some logic thinking in this discussion soon...

Posted by Joerg Moellenkamp in Privacy at 17:48

## **Takeoff 2**

Yes, i know ... a starting aircraft again. But this version is much better than the other one: You can see the thrust of the engines. I assume it's an A320 (position of the small fins on the aircraft, the look of the tailstrike protection and the form of the vertical tail).

Posted by Joerg Moellenkamp in Photographie at 17:02

## **Hoerempfehlung: Leigh Nash - All Along the Wall (Styrofoam Remix)**

Wieder mal eine Hörempfehlung für ein einzelnes Lied: Vielleicht erinnert sich noch jemand an das unsägliche "Kiss me" von Sixpence non the richer. War eine Zeitlang mein Lieblingsgrund das Radio wieder auszuschalten. Ein Song aus der Kategorie "Annoying Song - played damned too much".

Man darf sich aber fragen, was die Sängerin Leigh Nash heute macht. Denn sie macht richtig gute Musik. Insbesondere wenn Sie mit Arne Van Petegem zusammenarbeitet. Ich möchte hier allen "Kiss me" geschädigten All Along the Wall im Styrofoam Remix wärmstens ans Herz legen.

Posted by Joerg Moellenkamp in Music at 08:04

## **Nice gadget ...**

There is an factory installed barcode scanner in the the E71 (but you can install one on many mobiles with a camera). With this barcode you can encode your vCard in a way a camera phone is able to scan. So you can add it in a few seconds to your address book:

You will find the generator for such images at [mobilecodes.nokia.com](http://mobilecodes.nokia.com).

Posted by Joerg Moellenkamp in Fundsache at 07:16

Sunday, October 26. 2008

## **Takeoff**

Posted by Joerg Moellenkamp in Photographie at 18:03

## **Narbonne**

Posted by Joerg Moellenkamp in Photographie at 12:46

Friday, October 24. 2008

### **Nacktscanner**

Hmmm ... die Diskussion, die dieser Tage durch die Medien getrieben wird, kommt mir seltsam bekannt vor: In "Maximal invasive Sicherheit" habe ich 2005 erstmals ueber Backscatter-Röntgen geschrieben. Ich muss mittlerweile sagen, das ich weniger Probleme mit der Technik habe, als damaligerzeit. Mittlerweile sind die Geräte soweit entwickelt, das sie den Körper aus dem Bild herausfiltern und somit nur noch Fremdkörper auf dem Bild zu sehen ist. Also nicht die "Kleine-Jungs"-Phantasie mit der durch Klamotten guckenden Brille. In den Flughäefen in den USA ist es uebrigens ueblich, das die Bilder nicht von den Leute geprueft werden, die auch die Menschen dazu sehen, sondern diese in einem getrennten, entfernten Raum gescreened werden.

Posted by Joerg Moellenkamp in Aviation at 09:53

### **Bahnfahren**

Die Geschichte mit dem Kind, das von einer Fahrkartenkontrolleurin aus dem Zug geworfen worden ist, ist ja schon durch die Presse gegangen. Fefe berichtet jetzt in seinem Blog ueber weitere Einzelheiten:[...] stellt sich raus, dass die Bahn seit ein paar Monaten Geldprämien für erwischte Schwarzfahrer vergibt. Das würde natürlich erklären, wieso das Bahnpersonal das Kind rausgekantet hat, obwohl sie vorher noch einem Schwerbehinderten geholfen hat und der sie auf seinem Ticket kostenlos mitnehmen konnte und wollte.[.]Es bleibt zu hoffen, das eine Person mit so wenig Augenmass dem Arbeitsmarkt möglichst bald wieder zur Verfügung gestellt wird.

Das Kundenbindung nur sekundäres Ziel bei der Bahn ist, ist mir in letzter Zeit auch wieder häufiger aufgefallen. So hat mir eine Zugbegleiterin versucht zu erkläeren, das ein Aufzahlen einer Fahrkarte ohne IC-Zuschlag auf ein Ticket mit IC-Zuschlag im Zug nicht mehr moeglich ist. Das ist insbesondere dann dümmlich, wenn man am Abfahrtsbahnhof keine IC-Tickets mehr erhalten kann, weil man nach der Schliessung des Fahrkartenschalters auch dem Fahrkartenautomaten diese Funktion genommen hat. Begründung war: Die Software der Handheldterminals gibt das nicht her.

Man hat sich dann nach einem Streitgespraech darauf geeinigt, das Problem anders zu loesen. Es ist da vorteilhaft, wenn man noch vielzuvielen alte Fahrkarten in der Geldboerse hat.

Posted by Joerg Moellenkamp in Bahn at 06:28

Thursday, October 23, 2008

### **Andy will work for Arista Networks as the Chief Development Officer**

Hmm ... there is a lot of buzz around the leave of Andy. I was shocked in the first moment. But as usual nowadays, there are often missing parts in reports. I don't know what i'm allowed to say here, thus there will be only one cryptic comment from my side to this: Don't take all this stories at Forbes or Businessweek for granted. The real story is pretty much different from the stories in the most of papers so far. The NYT got it fairly right: But he said he would retain a part-time advisory role at the company. "It's my baby," Mr. Bechtolsheim said. "I will always be associated with Sun."

Posted by Joerg Moellenkamp in Sun at 11:05

Wednesday, October 22, 2008

### **HP Q3/2008**

The third quarter of HP is the one we call the fourth quarter at Sun. So it covers the time from April to June. There is one interesting chart in the earnings presentation of HP - the revenue for Enterprise Servers and Storage :

The next earnings announcement will be an interesting one. In the next earnings announcement they can't say that they grew year over year in this area. Nevertheless, there is a clear downward trend in this chart. Perhaps this is the reason why they bought EDS. They need that business to cover the decreasing revenue from their server branch.

Posted by Joerg Moellenkamp in The IT Business at 19:25

### **IBM results Q3/08**

There is a lot of interesting stuff in the 3Q 2008 Earnings Presentation of IBM (you find the pdf here). Let's have a look at page 12: Legacy System i is 82% down, i assume the 7% up for converged System p comes from customers buying pSeries instead of iSeries. xSeries is down 18% and storage 3%. IBM Microelectronics is down 27% year to year. In total the System&Technology business lost 10% (at constant currency it's even 11%).

So you'd easily say: Software and Services rescued IBM in the last quarter. But not everything is looking bright there. The signings for Global Services for short-term is up, but the signings for long term deals decreased 19% for strategic outsourcing, 17% down for Global Technology Service and 16% down for Global business services.

By the way: SMB is mentioned often as the big strength of IBM. They make 4,7 billion \$ dollars out of this segment. But you have to take into consideration, that IBM has a strong retail equipment department (point-of-sales terminals for example).

Posted by Joerg Moellenkamp in The IT Business at 15:25

### **New blades from Sun**

Sun announced a number of new blades for the Blade 6000 chassis yesterday. From my perspective two of them were especially interesting. The first one is the Sun Blade 6000 disk module. It's an 8-slot disk module based on SAS (you will find in-depth information in the Sun Blade™ 6000 Disk Module Configuration Guide manual).

The other important was the Sun Blade T6340 server module. This blade is equipped with two UltraSPARC T2+ procs. Thus you have 128 threads in one blade in a system and up to 256 GB of memory. When you look at the blade, you will recognize that a large part of the circuit board is occupied by memory slots. At the moment this is the blade with the biggest memory capacity in the industry.

In addition to this both blades we've announced a dual-proc Opteron based blade with the X6240 for the Blade Server 6000 chassis. The Blade Sun Netra CP3250 ATCA. This isn't a standard blade for the 6000 or 8000 blade chassis. It's a Xeon server blade for usage in telco blade servers.

Posted by Joerg Moellenkamp in Sun at 13:00

Tuesday, October 21. 2008

### **Outright depressing ....**

The good thing ... we´ve still more than enough money to get through the financial crisis and i wouldn´t wonder, if we still have positive cash flow .... (besides of effects like stock buy backs)

Posted by Joerg Moellenkamp in General at 21:07

### **1000 subscribers**

Posted by Joerg Moellenkamp in About this blog at 14:58

### **links for 2008-10-21**

Digg - Who dugg or blogged: Is the Linux community afraid of Opensolaris?

(tags: digg)

Posted by del.icio.us in del.icio.us at 13:00

Monday, October 20. 2008

## **Cheap benchmarking trick**

I already wrote about the latest benchmarking trick of IBM in the last blog article, but this article was in german, so i repeat this in english. There was an bold statement in a recent press announcement of IBM - "IBM Builds on Industry-Leading UNIX Portfolio With New Servers, Software":

The Power 560 can save up companies up to \$840,000 and 80-percent in energy by consolidating 13 Sun Fire V490 servers on a single Power 560 server with PowerVM, as compared to consolidating the same number on four Sun SPARC Enterprise M5000 servers with dynamic system domains."I've asked myself, how they get to such numbers. This number of servers couldn't based on performance. We don't need 4 M5000 just to substitute 13 V490. But after thinking about after reading the article in the Computerwoche i found out what's the trick of this comparison. The trick is a cheap one ... even for IBM marketing.

You can partition an M5000 in up to 4 domains. When you just want to consolidate 13 servers, you obviously need 4 systems. This comparison doesn't compare the compute power of the M5000 with the compute power of the p560. It compares two different virtualisation technologies. So the even the choice of 13 V490 is a really perfidious one. Twelve systems to consolidate would lead to 3 M5000, 13 systems lead to 4 because you have one domain too few. But that's not the point: You won't consolidate 13 V490 by using domains. You would use Solaris Containers (perhaps in conjunction with Solaris 9 Containers) for this tasks. By using this Containers you would need only one system, too. And you would need less processing power for it, as Container are a more efficient virtualisation technology in comparison to \*PARS.

By the way: The answer "one system" is false for both systems. Independently from the system architecture, virtualisation technology you want at least two systems and a cluster when you consolidate 13 systems on one. Without an additional standby system you are toast in the case of a system failure or maintainance. But that's a persistent error in every benchmarking comparison of IBM.

Posted by Joerg Moellenkamp in The IT Business at 17:14

## **Liebe IBM, das ist nun echt nen selten bloeder Trick ....**

Ein Kollege hat es schon ganz richtig gesagt: Dieser Artikel in der Computerwoche ist eine Denial-of-Service-Attacke auf das Sun-Marketing - IBM zündet Server-Feuerwerk. Denn was da erzaehlt wird, ist auch fuer IBM-Marketing-Verhältnisse echt billig. So billig, das selbst ich ein paar Minuten brauchte, um dahinter zu kommen, was es damit auf sich hat (und ich beschäftige mich häufiger damit) ...

Der "IBM Power 560 Express"-Server verbindet laut Hersteller die Power6-Prozessor-Technologie mit verbesserter Virtualisierungsfähigkeit und Energieeffizienz. Er eigne sich insbesondere als Server-Konsolidierungs-Plattform oder als Datenbank- oder Applikations-Server. Wie der IT-Riese erklärt, können nun Mittelständler durch die Konsolidierung von dreizehn "Sun Fire V490"-Servern auf einem einzigen Power-560-Server mit PowerVM-Technologie bis zu 80 Prozent der Energie sparen im Vergleich zu einer eventuellen Konsolidierung derselben Server auf vier "Sun Sparc Enterprise M5000"-Servern mit Dynamic System Domains.Ich hätte mir hier von der Computerwoche ein wenig mehr Recherchetätigkeit erwartet, als bloss eine Pressemitteilung von IBM abzuschreiben.

1. Wird hier wieder der uralte Benchmarking-Trick benutzt eine eine V490(eine Maschine die sich gerade in der Abkündigung befindet) gegen eine gerade angekuendige Maschine zu vergleichen.
2. Ich weiss uebrigens nicht, ob man das als Serverfeuerwerk bezeichnen sollen... eine p560 ist lediglich eine p570, die in ihrem Ausbaupotential halbiert worden ist. Also nix Neues.
3. Rechnen wir doch mal durch, wie man auf vier M5000 gekommen ist. Es geht hier um die Ablöse von 13 Rechnern. Jetzt möchte die IBM diese durch 4 M5000 abloesen. Wie kommt man auf eine solche Anzahl? Es liegt nicht an der Rechenpower, sondern an diesem kleinen Kommentar wegen der Partitionierung. Das ist recht einfach. Eine M5000 laesst sich in 4 Partitionen aufteilen. Nutzt man diese als Partitionierungstechnik, so braucht man natuerlich zur Darstellung von 13 Rechnern 4 Systeme. So ist auch die Anzahl 13 in einer aeusserst perfiden Art gewaehlt. Bei 12 Systemen, kaeme man schon mit 3 Systemen aus, wenn man denn so konsolidieren würde. Das ist eben diese Utilisation-Quatsch (20\$ fuer Sun, 60% fuer IBM) nur etwas versteckter.

Dabei muss man anmerken: Diese M5000 sind weit davon ab, auch nur ansatzweise ausgelastet zu sein. Man braucht sie nur wegen der Limitierung der Domains auf 4 pro System. Das ist der ganze Trick dahinter. Macht man das mit Containern - also auf Betriebssystemebene - kommt man auch hier locker mit einem System aus. Wobei Sun wahrscheinlich dann sogar noch mit einer weniger ausgebauten Maschine auskommen wuerde, da wir den Overhead der Virtualisierung via \*PARS nicht haben.

Sehr interessant ist auch, das man diese seitens IBM immer noch als Dynamic System Domains bezeichnet. Die hiessen so zu Sun Fire E10 und E20k-Zeiten. Man hat wohl bei IBM noch nicht mitbekommen, das sich die Zeit bei der Konkurrenz auch weitergedreht hat.

Und was hat es jetzt mit dem Kommentar der Denial-of-Service Attacke auf unser Marketing auf sich? Naja, jedes mal, wenn wieder ein solcher Bullshit durch unsere Presselandschaft lanciert wird, muss sich unser Marketing wieder darum kuetzen so einen Mist zu kommentieren.

Das Dumme ist: Man kann IBM dafuer nicht mal für vors Schienbein treten, den technisch ist das durchaus richtig, man kann eine andere Firma nicht für eine Darstellung einer dümmlichen Architektur verklagen.

Posted by Joerg Moellenkamp in The IT Business at 16:22

### **links for 2008-10-20**

Sun Academic Initiative

(tags: sun social education learning java students sai)

The Power of Xargs - Chris's Corner

(tags: unix sysadmin ssh scripting xargs solaris)

Posted by del.icio.us in del.icio.us at 13:00

Sunday, October 19, 2008

## **Is the Linux community afraid of Opensolaris?**

Okay, okay ... i know the headline is a little bit provoking. But when you think about some comments from Linux proponents you could think so. In the last few weeks i've heard one sentence quite often: "Why you you still develop Solaris? You should contribute to Linux!" from people administering Linux systems. And you could read at other places, that Solaris is irrelevant, that there is nothing worth of mentioning it or even for an integration to Linux. Just think about the Zemlin quotations! Or several other comments of proponents of Linux.

This is an interesting development. In the years before, there wasn't such comments. Solaris was considered as a dead end. But then the game changed. We open-sourced Solaris. The full monty over the time. We open-sourced the cluster framework. And we won't stop to open source further code until there is no more code to open-source. BTW: I find "Sun should contribute more" really interesting. In the moment you start up your text processor on your favourite Linux distribution you've gone through more code contributed by Sun than of anybody else. You've already traversed a large amount of code contributed by Sun when you just login into GNOME. This is a fact most people tend to ignore.

It's really interesting, that i find more and more articles that shows Solaris in a positive light. The reports on heise.de and especially the comments in forum are getting more positive for Sun and Solaris. In public forums you find more and more articles regarding Opensolaris. I would think about it as a good indicator for an increased uptake of Opensolaris in the market. Interestingly there are people who think about orchestration, when there is an increasing share of Solaris discussions on other forums, not believing that there is an increased interest in Opensolaris like this article.

And this is one of my important question: Why is there an increased amount of articles and comments disavowing the relevance of Solaris? Why are there no comments about BSD? Why are there no comments regarding AIX or HPUX? What's so special about Solaris? Or to ask the provocative question from the headline: Is the Linux community afraid of Opensolaris? Why is the Linux community afraid of Opensolaris?

At first: There is an easy answer to the question asked in the mentioned article regarding Mysql and Opensolaris and this fear of some people regarding Mysql and Linux. It's a matter of fact, that most Mysql installations run on Linux. There is more money to make with Mysql on Linux with Mysql on Solaris. And Sun has to make money. Just tell me one reason, why we should limit our money making oportunities by limiting Mysql to Solaris. And when there is a large effort for Mysql on Opensolaris, it's just because of leveling the playfield, as the development of Mysql was a linux centric one in the past (e.g. the scalability of Mysql is somewhat limited to the scalability of Linux respectively to the common sizes of systems used for Linux)

Of course ZFS is important for Solaris to stay relevant. Dtrace is important to stay relevant. An operating environment needs posterboy features to attract new users. An unix is an unix is an unix. Just doing another flavour of Unix isn't an incentive for someone to change to another unix. Of course other features are interesting as well, but they won't attract users. I once coined the phrase in a customer presentation: "There are cool features, and there are important features. There are features to attract new users. And there are features to keep users on your platform". Binary compatibility is a cool feature, but it's a feature you learn to love at the next migration to a new mayor release of your operating system. This is a feature that keeps you at Solaris, but it won't really attract you to change the platform.

By the way: Don't underestimate the importance of Dtrace. Of course it's not a feature for the every day user of Solaris. It doesn't have to be one. I find the lack of fantasy in the Linux community in regard the role of Dtrace a little bit astounding. When a developer recognizes the usefulness of dtrace, she or he will use Solaris, when she or he uses Solaris, more software will optimized or developed at all for Solaris (it begins with decent makefiles, goes to support for compiles other than gcc and ends with the usage of nice features of Solaris making the developers life easier). And more software running well on Solaris leads to a bigger user community and bigger mind share.

So, why does Sun doesn't contribute in large scale to the Linux kernel (we contribute a lot of code to Linux as an operating environment as i have stated before)? I'm sure it would help Linux. The Solaris engineering is one of the best or the best in the world of operating systems. But such a move wouldn't really help Sun. Linux and Solaris have pretty much different design principles. Linux and Solaris have pretty much different targets of systems. And the design principles of Solaris are imporant to Sun: Binary Compatibility. Scalability - a quadsocket Victoria Falls systems is essentially an 256-way SMP problem. Maintainability. How to update a running system while retaining a running version of a system. You can't sell high-end highavailable everything-redundant servers, without having an operating system that is capable of certain mechanisms that allow updates in the shortest possible time (and no, i don't consider the

in-situ updating process of Linux distributions as a good way to update packages. I really want Live Upgrade or Snap Upgrade before using it on a system in the M4000 and beyond range).

Even the development process is differently enough to justify the development of Solaris. It's the concept of the Architecture Review Comitees, the existence of a Solaris Sustaining Engineering. It's important to have such entities. Sun needs such entities to satisfy the needs of our customers. The customers expect patches for older releases, people just goaled to fix bugs and not to develop new features. It's one of the non-technical features people like. And we can't simply discard it, just to start developing at Linux. And i'm sure Linus and Linux doesn't want to take over the development concepts of Solaris.

You don't think, that's such an big issue: Even top contributors to Linux like Andrew Morton thinks, that Linux has it's own share of problems: (here(2005), here(2006),here(2007) and here(2008)). It's a reoccurring theme in every year. Problems that may lead to the need to restart the Linux franchise in the distant future. But that's a different story.

In my opinion, Linux is a good-enough Unix as x86 is a good-enough processor architecture. You can solve a good amount of problems. At some customers 100% (when you have hundred thousands of nodes, even good-enough is more than enough), at other the share-of-wallet for Linux is 0%. There is a reason, why we sell M9000/32 systems in good amounts despite of all sayings that you could solve all problems with a bunch of x86 servers ... harharhar. You shouldn't discard this need of computing in the area of "better than good enough" as a small niche. You have to consider, that both environments have pretty much different customers at their respective extremes. There are people using Linux on WLAN routers and there are people running SAP central instances on a full-blown SPARC Enterprise M9000.

Albeit the desktop experience of Solaris got better in the last versions, but it's a server operating system by it's DNA. Linux is something between a desktop and a small server operating system. Okay ... i'm sure somebody will say "Mainframe Linux! SGI Altix!" but you should really look at the architecture of this systems before taking them as an example for scalability. Linux must decide in the near future, what it wants to be. Without focus to the desktop, it will not be a valid alternative to Mac OS X. Without an focus to server operations it will never really get out of the small server sector.

I tend to look at the increased comments that we should stop to develop Solaris and start to contribute at Linux as a sign of people being nervous. They don't know, which role Solaris will play in the future. They assumed a waning relevance similar to other closed source operating systems like HPUNIX or AIX. But the opensourcing changed the rules of the game. When we announced Opensolaris, they said "Show us the code". We showed them the code. When we showed them the code, they said "Show us the community". We start to show them the community. Now they say "Is this community real?". There is a real community, there is real interest. And now we will see where this will end. I'm sure that this end wont be irrelevance.

And you should consider: When there were sure of their dominance, the Linux community wouldn't react in this way. They would ignore Solaris as they ignore \*BSD. When did you heard the last comment like "The BSD community should give up it's niche operating system and help us with Linux!" (Okay ... one reason for desisting from such a demand could be the consideration that one community is too small for two egos like Linus and Theo )

After all i find this comments "Hey, help us, don't develop an own system. Help us." from some parts of the OSS community a little bit strange. It's not the first time: When we announced the availability of an in-kernel CIFS stack, the same words came from the Samba community. Having alternatives is a good thing, at the end it's one of the reasons, why Linux got such an important building block of modern IT infrastructures. At which date "Having a choice" got a bad name ?

I'm not ridicilious. I do not believe that we will the a market share of Opensolaris larger than the one of Linux anytime soon. But i strongly believe, that we will see a market share that enables a vivid community around it. We already see the first startups using Opensolaris as the foundation of their appliances.

More important: Linux needs Opensolaris. Do you really think, that there would be any development fo5 brtfs without ZFS? I'm observing the development of Linux for quite a while (i'm using Linux longer than Solaris), but from my perspective the speed of the kernel development slowed down in the recent years. The Linux community would celebrate ext7 in 2015 (some new features per release, but essentially the same stuff). Linux needs a strong competitor for its further development as Solaris needed the wake-up call from the Linux community. The big step in functionality from Solaris 9 to Solaris 10 is in a part an answer to Linux.

And to comment the opening statement of this article at the end of my article: Not the purchase of Mysql will rescue the

Solaris franchise, it was Solaris 10 and Opensolaris that already rescued it.

Posted by Joerg Moellenkamp in Solaris at 21:31

### **Less known Solaris features: About crashes and cores - Appendix B: ::status**

You are logged into your system and doing a little bit of house keeping (archiving of old logfiles, deleting the rubbish on you system like backup files) and suddenly you see a core file. Heck ... you wasn't aware of the fact, that an application wrote such a file. How can you get some basic information about it? With Solaris you can use the mdb for this task.

Let's assume you find a core file at / on the system master: # uname -a

```
SunOS master 5.11 snv_97 i86pc i386 i86pc
```

```
# ls -l core
```

```
-rw-r--r-- 1 root root 5073385 Oct 10 18:26 core
```

```
Now we can start the Modular Debugger mdb with the core file.
```

```
# mdb core
```

```
Loading modules: [ libc.so.1 ld.so.1 ]
```

```
>After a short moment the mdb command will give you a prompt. Now just type in the ::status command> ::status
```

```
debugging core file of sshd (32-bit) from master
```

```
file: /usr/lib/ssh/sshd
```

```
initial argv: /usr/lib/ssh/sshd
```

```
threading model: native threads
```

```
status: process core file generated with gcore(1)Now you know, that you can safely ignore the file, as the system created it on your order while you was playing around with gcore.
```

Posted by Joerg Moellenkamp in Solaris at 11:30

Saturday, October 18. 2008

## **Nokia Sports Tracker Beta**

One of the nice features of my new Nokia E71 mobile is the integrated GPS receiver. So i searched for a tool to use it as a GPS tracking device for geotagging of photos. I found such a tool ... it's even from Nokia. This tool was developed for tracking sport workouts, but it allows you to export your tracks as GPX or KML files as well, so you can use it for photo geotagging (for people like me).

Yesterday evening i've decided to walk from cinema to home instead of using the subway and played around with the tool. You can directly upload at home. You can even activate "Life Tracking" so other people can look at your workout while you are at workout. I didn't use this feature. I'm not a fan of this kind of exhibitionism and obviously it's a security risk. It tracks speed, altitude, position and allows you to look at some analytic screens while tracking (for example time/speed, time/altitude and similar graphs)

You can share your tracks with friends or with anyone. I opened up the track from the picture above for everyone, so you can take a look at the data collected by the tool. The application for your mobile and additional informations are available at the website of the Nokia Research Center

Posted by Joerg Moellenkamp in General at 16:17

## **links for 2008-10-18**

[Phoronix] OpenSolaris 2008.11 Starts Coming Together

(tags: Opensolaris 2008.11)

Posted by del.icio.us in del.icio.us at 13:00

## **"Linux feels like it was written. Solaris feels like it was designed."**

Neil A. Wilson wrote an interesting article why he chosen Solaris as is favourite operating environment: "Why i like Solaris". It's an article about some features of Solaris, but there is one paragraphy that summarizes an important reason of my choice for Solaris many years ago as well:Linux feels like it was written. Solaris feels like it was designed.. While I think that Sun's development processes can sometimes be a little heavyweight, and I think that Sun is trying to retain too much control over OpenSolaris, there is a lot to be said for having processes in place to guide development.This is the basic difference. Linux is based on the concept "Okay, you can develop a feature and we will see if we integrate it into the mainline code." Opensolaris works differently in this area. The concept of PSARC may look as an heavyweight process for an opensource operating system, but it's the reason for "Solaris feels like it was designed." as the documented review of the design and the analysis of the impact to the whole architecture is the beginning of all code in Solaris.

Posted by Joerg Moellenkamp in Solaris at 11:46

## **Mitten in der Stadt**

Ich wohne ja nun wirklich so gut wie mitten im Hamburger Stadtgebiet. Hamburg is ja nun eine sehr gruene Stadt und ich habe auch noch das Glück in der Nähe eines Parks zu wohnen. Aber im Teich des Hammer Parks einen Graureiher zu sehen, hat mich dann doch ein wenig ueberrascht:

Sorry für die schlechte Bildqualität, aber ich hatte leider zu wenig Licht, zu wenig Zoom und noch keinen Kaffee

Posted by Joerg Moellenkamp in Photographie at 10:25

Friday, October 17. 2008

## **Project Sirius**

At last: The IBM mainframes will get a decent operating system. Project Sirius is at least one to two years away from prime time but at least it's a silver lining on the horizon.

Posted by Joerg Moellenkamp in Solaris at 16:08

## **links for 2008-10-17**

Guest View: Java + multicore = good news - SD Times On The Web

(tags: programming parallel development java)

Posted by del.icio.us at 13:00

## **Sieben Uhr morgens ... in Deutschland ... am Bahnhof**

Man mag sich ja fragen, ob die Zugehörigkeit eines Menschen in eine bestimmte Alterskohorte dafür sorgt, das man einen besonders wirren Klamottengeschmack entwickelt.

Ich will mich da nicht ausnehmen: Ich hielt es fuer eine gute Idee, eine zeitlang zu Abizeiten einen Trenchcoat zu tragen. Heute würde ich dafür wohl von der Schule verwiesen werden, da man kurzfristig damit gerechnet haette, das ich eine abgesägte Flinte mit in die Schule bringe. Aber zu meiner Zeit gab es noch kein Counterstrike, kein Halfife und man hatte noch sowas wie eine realistische Chance fuer die Zukunft.

Ich habe noch andere Todsünden begangen, und begehe sie bis heute ... oftmals weil mir einfach vollkommen egal ist, was gerade Mode ist, wie man etwas zu tragen hat. Mein Lieblingspullover ist auch heute ein quietschegelber (das Schild hielt es fuer mangogelb) Kaputzenpullover. Mit uebergezogener Kaputze sehe ich darin aus wie ein Teletubbie. Aber er ist warm, er ist gemuehtlich und ich passe auch noch nach zwei Jahren Frustfuettern rein.

Egal ... ich sitze hier im Zug (und gemessen daran, das mein Kaffeebecher von einer Seite des Tischs zur anderen wandert, scheint es der Kutscher sehr eilig zu haben) und gerade ist mir eines dieser Raetsel ueber den Weg gelaufen: Pseudouniform (Schulterklappen mit Goldrand) mit aufgestickten Initialen. Das sich rosa Hemden fuer Männer nicht mehr aus dem kolletiven modischen Missempfinden streichen lassen, ist mir ja auch schon klar geworden. Aber der Zirkusdirektor-Look ... das geht zu weit. Wenn das zum allgemein gültigen Look in Deutschland wird, werde ich mich entgütig um einen Job als Manager in Zentralsibirien bewerben.

Eine weitere interessante Beobachtung konnte man uebrigens heute morgen machen: Das man uebrigens nur Maenner (oder entsprechend im Larvenstadium: Jungs) stockbetrunken morgens am Bahnhof findet, ist uebrigens auch eine oft geaeusserte, aber ebenso falsche Mutmassung: Das Maedel, das so gegen sieben ausladenden Schrittes durch die Wandelhalle getorkelt ist, duerfte sich im Laufe des Tages ueber die Segnungen von Paracetamol, Ibuprofen et al. freuen. Um hier wieder den Kreis zur Mode zu schliessen: Teuer in der Kleidung, billig im Geschmack ist auch hier problemlos moeglich. Gut, das Frauen so um die 20 nicht mehr in mein Beuteschema passen ... ich glaube ich hätte sonst ernste Probleme mit der Paris-Hiltonisierung insbesondere der weiblichen Anteile der Gesellschaft.

Wenn dann auch noch die von einer Bekannten geaeusserte Vermutung eintritt, das sich in den USA (merke: Sex in the City und New York hat soviel mit den USA zu tun wie Batman mit Polizeiarbeit) Frauen eigentlich sich die Kleidungswahl auf die Auswahl eines Rosa-Farbtons beschaerakt (jun g equals schreiend rosa, alt leichtes rosa, fast weiss) und das vielleicht nach Deutschland rueberschwappt, dann faengt sich mein Magen schon heute zu kraeuseln an. Schwul werden ist uebrigens keine Alternative ... bis dahin hat sich der Circusdirektor-Look fuer alle durchgesetzt.

Posted by Joerg Moellenkamp in Bahn at 09:16

Thursday, October 16. 2008

### **The Register about future SPARC developments**

TPM wrote an article about some comments of Sun and Fujitsu executives and enriched it with some speculations - Sun and Fujitsu hint at Sparc futures:As part of the launch of the Sparc T5440 midrange server this week in San Francisco, top brass from both Sun Microsystems and Fujitsu spent some time assuring customers that the companies' chip and systems partnership going strong and that both were working away on Sparc processors that would end up in future systems.

Posted by Joerg Moellenkamp in General at 20:26

### **Heise.de über die Sun SPARC Enterprise T5440**

heise.de hat einen sehr wohlwollenden Artikel über unsere neuen Server auf Basis des UltraSPARC T2+ prozessors veröffentlicht: Suns Enterprise-Server T5440 mischt die Szene auf:Mit dem neuen Vierprozessor-Server SPARC Enterprise T5440 stellen Sun und Fujitsu unter Beweis, dass mit der SPARC-Architektur in der Serverwelt weiterhin zu rechnen ist.undMit solchen Ergebnissen haben Sun und Fujitsu eine gute Chance, verloren gegangenes Terrain im HPC- und Supercomputer-Bereich wieder zurück zu erobern.

Posted by Joerg Moellenkamp in General at 13:23

### **links for 2008-10-16**

BigAdmin Description - An Open Source Web Solution - Lighttpd Web Server and Chip Multithreading Technology

(tags: cmt sun solaris lighttpd)

Posted by del.icio.us in del.icio.us at 13:00

Wednesday, October 15, 2008

## **Analysing a so-called "Comparison" about Virtualisation at IBM Developerworks**

Whenever you want to dismiss the claims of a competitor or want to set your own or preferred technology in a better light, you should do some research on your topic. Otherwise you may end up with a document that's outright ridiculous.

I found a really strange piece of "comparison". It's called "A comparison of virtualization features of HP-UX, Solaris, and AIX". It's written by Mr. Ken Milberg. And I wasn't able to stop my shaking the head in disbelief. This text reinforces my personal impression, that this author is just a hired gun to publish claims even IBM doesn't want to make. But let's dissect his newest blurb. You shouldn't read it ... it's just a really abysmal document. I've sacrificed my time to do it for you, so don't waste your own

I will start with just four sentences of Mr. Milberg's document:

Scalability -- Only eight CPUs and 64GB RAM on one machine

Server-line -- Only low-end Sparc servers are supported

Limited micro-partitioning -- Four partitions on T1, Eight on T2

No Dynamic allocation between partitions

The truth is:  
Scalability: 4 sockets, 32 cores, 64 pipelines, 256 threads, 512 GB of memory

Server-Line: I would call the Sun SPARC Enterprise T5440 not really low-end SPARC services

Micro-Partitioning: Up to 128 LDOMs on a T2+ system, Up to 64 on a T2 and 32 on a T1.

Of course you can resize the LDOM without rebooting the system. At the end this is the way, you initially configure the system: At the beginning all CPUs and memory resources belong to the control domain, you take them away from this domain to give them to the guests.

Four claims, four times utter bullshit. I could stop the dissection now, but the article starts to get even more funny. So let's go ahead: Sun also offers hardware partitioning, which allows their high-end servers to be divided into four-process partitions. These are referred to as Sun DSD's. In many ways this technology is similar to IBM logical partitioning, which was introduced in 2001, with no real virtualization capabilities. DSD was the name of the technology in the Sun Fire Enterprise line. You were able to split those systems at the granularity of system board, thus the granularity was 4 sockets per domain minimum. That's correct.

But the Sun Fire Enterprise almost reached the end of its lifetime and now we sell the M-Class systems. There is something called quad-XSB mode. The documentation for Dynamic Reconfiguration on the M4000/M5000/M8000/M9000 states: SPARC Enterprise M4000/M5000/M8000/M9000 servers have a unique partitioning feature that can divide one physical system board (PSB) into one logical board (undivided status) or four logical boards. A PSB that is logically divided into one board (undivided status) is called a Uni-XSB, whereas a PSB that is logically divided into four boards is called a Quad-XSB. Each composition of physical unit of the divided PSB is called an XtendedSystemBoard(XSB). These XSBs can be combined freely to create domains. In a M4000 you can create 2 partitions, in a M9000 you can create up to 24.

Despite the statements of Mr. Milberg you are able dynamically move resources from one domain to another. This is a really old trick. I've done this on one of my E10K in 2000 and the system with the capability of creating 16 DSD. And this system was introduced in March 1997. So Mr. Milberg's comment "... IBM logical partitioning, which was introduced in 2001 ..." is good for some amusement.

At the end of the Sun part of the document he even starts to celebrate the advantages of WPARs without mentioning the disadvantages.

To close my article: The whole article is an insult to the real meaning of the sentence "This article explores all of these topics in detail." I'm not really sure how such an article was able to pass the editorial quality control at IBM.

PS: Okay, when you really want to read this botch job ... here is the URL: A comparison of virtualization features of HP-UX, Solaris, and AIX

Posted by Joerg Moellenkamp in The IT Business at 22:46

**Is iSCSI really dead? The Register thinks so ...**

The Register thinks that FibreChanneloverEthernet(FCoE) will take over the datacenter and proclaims "iSCSI: Game over". Okay, no problem .... there is already a FCoE implementation for Solaris. Been there - done that. There is only one simple problem: I'm not sure, that FCoE is the way to go.

At first: You don't get really an advantage by FCoE in relation to iSCSI. Using FCoE for storage connectivity means: SCSI commands over Fibrechannel over Ethernet. iSCSI is SCSI commands over IP over Ethernet. Same number of layers. So nothing to win here. And it doesn't look as way to a single fabric datacenter. I don't think that even a more evolved and specialized Ethernet can solve the problems of increasing latencies when you use it for several services at the same time. What's more important writing the data to the client or getting the data from the disk? But let's assume, the latency issues are solved in an acceptable manner.

In return of using FCoE, you loose proven technologies: You can't route between networks, you can't use IPsec for encryption and authentication and integrity checking. In IP you have already a proven wide-area naming service (you call it DNS) and you have your IP networks available today. In iSCSI you have already an established naming system called iSNS. With hexa- and octacore processors you have more than enough clock cycles for the TCP/IP stack.

There is an important fact: The E in FCoE isn't the Ethernet of today. You can't simply use your existing Ethernet-Network. FCoE needs an enhanced Ethernet. It isn't usable on normal Ethernet as Ethernet is a lossy transport by nature. Higher protocols in the TCP/IP stack solve this problem for the application at the moment. But the Ethernet of FCoE has to do this on its own. So you need switches capable of providing a lossless Ethernet to use them as your storage connect.

iSCSI simply uses TCP/IP to provide a lossless communication, thus you could even use a wide area network for a storage interconnects. Try this with FCoE) .The TCP/IP stack and adjacent services (routing,naming,encryption,authentication) are a proven piece of software in most operating systems with a common set of standardized protocols with their share of time to iron out incompatibilities. Why should we develop and debug all this stuff again?

The whole FCoE talk looks like the last stand for the FC vendors to fight for their market share before storage networks are migrated to IP as well. And when you really want to establish a decent new storage interconnect: Use Infiniband as the foundation. 40 GBit/s (so you have more than enough headroom for multiple services in parallel) and really low latencies. It gives you iSCSI as a well established standard and RDMA over Infiniband to take TCP/IP out of the equation.

Posted by Joerg Moellenkamp in Sun at 18:49

## **links for 2008-10-15**

US Air Force Saves money/space with Solaris containers : Jim Laurent's Weblog  
Are Solaris containers "certified" for use by the US Government or DoD?

\* Short answer: Yes! Read on for the long answer.  
(tags: container dod government)

Sun Ray Software 4 10/08 - Released! - Think Thin

(tags: SunRay)

Posted by del.icio.us in del.icio.us at 13:00

## **John Cleese: Michael Palin isn't the funniest Palin anymore**

(via: Egghat)

Posted by Joerg Moellenkamp in General at 08:33

## **Credit default swaps - oder wie kamen die Riesensummen zustande ...**

## **Blog Export: c0t0d0s0.org, <http://www.c0t0d0s0.org/>**

Ein Kommentar beim Rebellmarkt erklart, wie diese absurden Summen an ausstehenden CDS zustande kommen: Die Magie der CDS. Dieser Kommentar erklart den Irrsinn ziemlich gut.

Posted by Joerg Moellenkamp in General at 07:51

Tuesday, October 14. 2008

### **Nokia E71 - or: Small features ...**

Yesterday i've got a new mobile. It's one of this superduperallinclusive smartphone. There is even a GPS receiver in the device. Additionally the phone includes all this hip three to six letter acronym technologies. But the best feature of the mobile doesn't need all this arcane technological wizardry.

Before tell you about this feature, i have to describe one of my habits: I use my mobile as a alarm clock at home as the ringing of a telephone is more efficient than the ringing of a alarm clock as a high priority sleep interrupt. I assume ringing of the telephone is hardwired to "an important message" whereas the sound of an alarm clock is hardwired to "waking up and leaving the warm bed" in my brain.

Okay ... the absolute best feature of the E71 is: When you press the middle of the cursor button of the locked mobile , it displays the time white on black (so it doesn't dazzle you) with a font size you can even read with the small eyes at 05:00 o'clock when you have to reach the red-eye train ...

Posted by Joerg Moellenkamp in Technology at 21:40

### **About going private ... and about gambling**

TPM writes in "Sun, Novell, and Cray - Time to go private?":The company at the very top of the list is - no drum roll needed - Sun Microsystems. While Sun's Solaris is one of the best operating systems on the planet and its servers are decently engineered, the company is spending far too much time justifying its strategies to investors and not enough time selling its products and building up its channelsThis is a really rare moment ... but i'm sharing the opinion of TPM. My very personal opinion is, that he is correct: Sun should go private. Our product portfolio is excellent at the moment, i'm aware of many future developments (but i'm not allowed to talk about them ... in short: You Ain't Seen Nothing Yet). I'm perfectly sure that our strategy will pay off, but this will need some additional time.

But you can't do this job with a bloodhound gang of analysts in your neck and headcount reductions to appease the financial community. And obviously we have the best operating system on the planet

PS: By the way - sometimes the financial markets looks more like a professional version of Las Vegas to me ... could you explain the existence of more credit default swaps than credits otherwise? As far as i remember the whole stocks thing was invented to find investors ... but this time must have been passed a long time ago.

Whenever investors were outnumbered by players (buying and selling stocks just based on the quotes and chart analysis without knowing much about the fundamentals of the company is not that different from playing Blackjack or Poker) we really had problems: The credit crunch now , the world economic crisis in the thirties of the last century, the NewEconomy bubble. Gamblers tend to panic in the case of unforeseen or large losses. Long time investors look at them as a chance (Evidence: The horde of investors versus Warren Buffet at Goldman Sachs)

Perhaps we should introduce a tax on stock trading: The shorter you hold a financial product, the more tax you pay on it. When you buy some stocks or papers to hold them for years (retirement funds or something like that) you would pay no taxes on the stocks. The more speculative a financial product is, the more tax you pay on it. Maybe such a instrument would give us a sane, a investors stock market back (at least we could pay the next "butt-saving-bailout" with the money)

Disclaimer: This is my private opinion. Really. A private opinion in the sense of: "Thinking about it in the evening while sitting in my living room with a glass of a good red wine ..."

Posted by Joerg Moellenkamp in Braindump at 20:17

### **A kind of Time Machine**

The Time Machine in Mac OS 10.5 saved my butt a few times in the last few months. Getting back an older version of a presentation when you've saved a short version of you presentation in the same file as the long version a few days earlier is quite handy. Solaris is now capable to do a similar thing (albeit the implementation is vastly different): Time

Slider is based on periodic ZFS snapshots and presents them in a nice way: Time slider is one of the new features that will be available in OpenSolaris 2008.11. Time slider provides an automatic way to backup your data on the same disc using one of Sun's ZFS filesystem unique features, snapshots. With time slider you can browse and recover files from snapshot backups using the GNOME file manager. When you click on the life buoy in the File Browser you get a slider to step back in time based on this snapshot ... wait ... just look at the screenshots at Erwann Chénéde's weblog. A picture substitutes 1000 words and this feature is outright incredible. Excellent work!

Posted by Joerg Moellenkamp in Solaris at 19:58

### **links for 2008-10-14**

Turbulence Forecast - Atlantic Ocean Westbound Tracks

(tags: aviation)

Turbulence Forecast - Atlantic Ocean Eastbound Tracks

North Atlantic Tracks - Wikipedia, the free encyclopedia

(tags: navigation aviation)

Wie Marcel Reich-Ranicki im hohen Alter keinen Fernsehpreis, aber dafür eine Tochter bekam - Spreeblick

(tags: MRR medien)

Sozialtheoristen - Blog Archive - Vertrauenskrise? Die 45-Billionen-Dollar-Lüge

(tags: creditcrunch)

Meggison Technologies: Quoderat - Blog Archive - REST: the quick pitch

(tags: xml rest webservices)

Posted by del.icio.us in General at 13:00

### **Mysql and ZFS compression**

Don gave an update to his article about mysql on Opensolaris. Now he published an update to his article with ZFS & MySQL/InnoDB Compression Update: Conclusion? Unless you care a great deal about eking out every last byte (using a RAM disk, for example), LZJB seems like a much saner compression choice. Performance seems to improve, rather than degrade, and it doesn't hog your CPU. I'm switching my ZFS volume to LZJB right now (on-the-fly changes - woo!) and will copy all my data so it gets the new compression settings. I'll sacrifice some bytes, but that's ok - performance is king.

Posted by Joerg Moellenkamp in General at 06:48

### **My first article has gone live on SDN**

With a huge amount of help from Marina Sum the first article from the LKSF book made it to the Sun Developer Network: Introducing pfexec, a Convenient Utility in the OpenSolaris OS.

Posted by Joerg Moellenkamp in Solaris at 06:07

Monday, October 13. 2008

## **The architecture of the SPARC Enterprise T5440**

Denis Sheahan wrote two insightful articles to the architecture of the T5440. The first one is about the architecture of the system in general. The system is really an elegant design. By the way: The similarities to the X4600 chassis are not an accident

The second one is about a chip, that made this system possible at all - the Zambezi ASIC: Coming out of each UltraSPARC T2 Plus processor are 4 independent coherence planes. The T2 plus has 8 banks of L2 cache and each plane is responsible for the traffic from two of these banks. The plane is identified by two bits (12 and 13) of the Physical address. There are 4 Zambezi hubs in the system, each handling a single coherence plane. Each Zambezi is connected to each of the four T2 Plus processors over four separate point-to-point serial coherence links. Because planes are independent there are no connections between the Zambezi chips.

Posted by Joerg Moellenkamp in Sun at 20:44

## **L2ARC on ramdisks?**

I thought a little bit about the idea of transforming server into solid state disks. The idea in the mail of Chris Greer on zfs-discuss was to use mirrored iSCSI shared ramdisks as a storage for the separated ZILs. But i think you could use the concept as well for L2ARC as well - e.g. for large databases. One of the sizing rules of databases: More main memory never hurts. Nothing helps the performance of a database more than even more memory. The rule of "main memory never hurts" is based on the fact, that a hard disk has only a few IOPS compared with the main memory and hard drive access massively hurts the performance of your database.

But obviously the size of memory is limited, albeit the this limit is quite high with systems with memory sizes in the range of 512 GB on 4 rack units. But how can you get more memory into your database system, when all DIMM slots are filled with the biggest available DIMMS.

I had an idea while cooking tea this evening while i thought about a discussion with a colleague: Let's assume an architecture based on a X4600 as a head in front of four X4600 each fully maxed to 512GB. All the nodes are connected with Infiniband. The first X4600 is your normal database server (for example mysql or LarryBase). You put your data into an ZFS storage pool. This storage pool is augmented with L2ARC devices. But now comes the plot twist. Let's use the 512GB X4600 as huge ramdisks (yes, i know, every engineers heart will crying now) speaking via iSER (no TCP/IP, just RDMA) at 20 GBit/s to the central database node. This would give you a cache in the size of almost 2 TB plus the cache on the database server itself.. By using L2ARC you could use the memory as database caches of other systems without using a database doing a combination of the memory resources by other means, for example the CacheFusion stuff of Oracle. You don't have to fuse the caches of other databases servers. The other servers are caches. You don't have to partition the databases.

It would be interesting how such an system would perform in comparision to a Oracle RAC or other memory implementations. Anybody out there willing to test this ... my Infiniband switches are in the laundry at the moment

Posted by Joerg Moellenkamp in Solaris at 18:47

## **Compression**

Tim Cook wrote about some interesting observations in his blog: The Seduction of Single-Threaded Performance: Data compression utilities are a classic example of a seemingly mature area in computing. Lots of utilities, lots of different algorithms, a few options in some utilities, reasonable portability between operating systems, but one significant shortcoming - there is no commonly available utility that is multi-threaded.

Posted by Joerg Moellenkamp in General at 16:44

## **Vielen Dank!**

Ich möchte Herrn Reich-Ranicki meinen ehrlichen Dank ausdrücken. Es war nötig, das irgendwer der nur noch

selbstreferenziell tätigen Gruppe von Medienmachern mal zeigt, das es Menschen gibt, die sich nicht mit dem momentan Zustand der Medien einverstanden erklären. Es ist daran wahrlich wenig, das man feiern könnte und noch weniger, das einen Fernsehpreis verdienen würde. Auf Youtube ist der Auftritt von Herrn Reich-Ranicki bereits zu finden:

Posted by Joerg Moellenkamp in Braindump at 16:25

### **A really long mpstat output**

At 256 threads some of the Solaris commands have a really long output: Solaris on the T5440. Just look at the mpstat output at the end of the article

Posted by Joerg Moellenkamp in Sun at 15:56

### **Sun SPARC Enterprise T5440 announced**

256 compute threads, 512 GB of main memory and just announced : Sun (NASDAQ: JAVA) jointly announced the SPARC Enterprise T5440

"... 7,520 SAP SD Benchmark users ...", "14,000 active Siebel benchmark users was set on a single SPARC Enterprise T5440 server" and "Sun SPARC Enterprise T5440 (4 chips, 32 cores, 256 threads): SPECint\_rate2006 - 301, SPECfp\_rate2006" are just a few numbers you will find in the press release at Sun Sets Multiple World Records with New Solaris-Powered Sun SPARC Enterprise T5440 Server. You will find further informations about the system on the sun.com website.

The following numbers are really interesting as well. I did some calculations in the Sun SE T5440 power calculator. At full load fully loaded with options (512 GB memory, 4 HDD, 4 Procs@1,4Ghz, 2 PCIe cards, 2 XAUIs) the system just takes 1906 Watts. At full load but with minimal 4proc configuration (32 GB Memory, 2 HDD, 4 Procs@1,2 GHz, no XAUI or PCI cards) the system just takes 942 Watts.

(Substantiation of the benchmarks in the press release)

Posted by Joerg Moellenkamp in Sun at 15:23

### **links for 2008-10-13**

Make-Believe Maverick : Rolling Stone

(tags: usa toread politics news election article history)

Posted by del.icio.us in del.icio.us at 13:00

### **Anatomy of an attack**

Paul Murphy wrote an excellent article about the anti-Solaris article that even found it's syndication at the NYT. Paul writes in Anatomy of an attack: The New York Times on Solaris: The piece itself illustrates the standard recipe for attack journalism: use a title sure to attract the attention of editors sympathetic to your cause; pretend to balance in the article but actually use strong negatives and weak positives; find one or two third parties to attribute the really nasty stuff to; and, strip any facts you need of their real context while leveraging reader assumptions to add an emotional patina of your own.

Thus using "Sun Solaris" in the title instead instead of just "Solaris" or "OpenSolaris" invokes one of the not so secret handshakes characteristic of the anti-Sun community to grab the attention of any editor with a Microsoft or IBM agenda. It's an really interesting analysis of the article. I really thing that we will see more attacks in the future as Opensolaris get a more and more viable alternative to Linux in many "linux-only" shops. Paul ask at the end of it's article. "Qui bono". I don't know ... wait ... of course i know it, but i can't pinpoint the exact source. But i'm sure of more attacks of this kind. There is a market share to loose.

Posted by Joerg Moellenkamp in Solaris at 09:07

### **First media coverage about the Sun Fire T5440**

The Sun Fire T5440 isn't announced so far, but there is already some media coverage about it. The Computerworld cites in Sun doubles processing power of UltraSparc T2 Plus servers: Jean Bozman, an analyst at IDC, said that although Sun's overall server revenue has dropped off, sales of systems based on the multithreading UltraSparc chips "have seen dramatic growth." Bozman also said that she is seeing evidence of Sun gaining new customers via the multithreading technology. By the way: Jeremy Barnish did a nice interview with two guys of the Sun Fire T5440 development team: "Sun's Batoka Launch - "The Way of the Future".

Posted by Joerg Moellenkamp in General at 08:22

Sunday, October 12. 2008

### **links for 2008-10-12**

YouTube - Toy Story Requiem

(tags: youtube mashup toystory requiem)

WSO2 Web Services Framework for PHP | WSO2 Oxygen Tank

(tags: webservices rest restful)

Writing a Simple REST and SOAP Service With PHP | Dimuthu's Blog

(tags: webservice rest)

Posted by del.icio.us in del.icio.us at 13:00

### **The dark toy**

Posted by Joerg Moellenkamp in General at 10:47

### **Opensolaris at Smugmug**

Don MacAskill (the CEO of SmugMug - yeah CEO and geekdom isn't mutually exclusive) did an interesting experiment - he used an Opensolaris system for one of his database replicas. And his experiences were really positive (besides of the usual pet peeve we already working on ): I'm a Linux geek, have been since 1993 (Slackware!). All of SmugMug's datacenters (and our EC2 images) are built on Linux. But the current state of filesystems on Linux is awful, and it's been awful for at least 8 years. As a result, we've put our first OpenSolaris box into production at SmugMug and I've been pleasantly surprised with the performance. The configuration has a really interesting speciality. He used compression. Lo and behold, it worked! We're getting a 2.12X compression ratio on our DB, and performance is keeping up just fine. I ran some quick performance tests on large linear reads/writes and we were measuring 45.6MB/s sustained uncompression and 39MB/s sustained compression on a single-threaded app on an Opteron CPU.

Posted by Joerg Moellenkamp in Solaris at 10:09

Saturday, October 11. 2008

## **Solaris and the GNU (tools)**

The ever reoccurring customer criticism in the Q&A part of my presentations is: "Your command line tools are really strange in regard of the options. Why does your ps has different options than the Linux ps. By the way: You didn't cleaned up your operating environment as i found multiple versions of tool xyz. But i still didn't found any GNU tools. I have to reinstall every tool on my own".

Well ... i want to talk about this criticism. It comes down to the following questions: Why are the Solaris tools different to GNU. Why are there several versions of the same command in Solaris? Where are the GNU tools? The behaviour of our tool is differently to the GNU tools for examples. The standard toolset of Solaris is the SysV toolset. Of course you will find more binaries in /usr/bin and /usr/sbin than just the stuff from SysV but the basic tools (ps for example are SysV ones) are implemented with the SysV guidelines in mind. Why do we keep this tools? Why didn't we subsituted them by GNU variants? Simple answer: Binary Compatilby Guarantee. There may be programs out there dependent on a toolset with the same behaviour (regarding options and output) than in older versions of Solaris. One of the advantages in Solaris is the fact, that you can take a script from Solaris 7 and it will run without problems on an Opensolaris script. No changes to the script because of a different output layout or an renamed option. Our customers expect this. It's one of our selling points. So you can't simply change the behaviour of an existing tool.

But why do we have several versions of some tools? The reason for this duplication is again really simple: We follow standards. Yes, there are official standard bodies in the Unix world Sometimes the behaviour of a tool in a certain standard is different than in another standard. The man page states  
If the behavior required by POSIX.2, POSIX.2a, XPG4, SUS, or SUSv2 conflicts with historical Solaris utility behavior, the original Solaris version of the utility is unchanged; a new version that is standard-conforming has been provided in /usr/xpg4/bin. If the behavior required by POSIX.1-2001 or SUSv3 conflicts with historical Solaris utility behavior, a new version that is standard-conforming has been provided in /usr/xpg4/bin or in /usr/xpg6/bin. If the behavior required by POSIX.1-2001 or SUSv3 conflicts with POSIX.2, POSIX.2a, SUS, or SUSv2, a new version that is SUSv3 standard-conforming has been provided in /usr/xpg6/bin. When an application mandates the conformance of your system to a certain standard, you just have to set the PATH accordingly. The man page standards(5) describes the exact PATH to enable the conformance to a certain standard.

In addition to the paths with the commands needed to conform with standards, there is a directory with commands called /usr/ucb. UCB? University of California Berkeley. The B in BSD. When you prefeere the BSD-style of commands (good example is ps auxww instead of ps -ef) this directory will provide you those commands. The existence of this commands is largely a legacy from the time before the BSD to SysV conversion. Until 4.x the operating System was a BSD with some SysV concepts included, 5.x (better known as Solaris 2.x) is a SysV with some BSD concepts. In early years the dispute between BSD and SysV was a holy war equal to the one about the best editor. Nowadays the line between both camps gets more and more blurry.

Okay, The GNU tools do not really follow any standards. So the commands in the various parts of the system doesn't help you. They work differently, some even implement SysV and BSD command options and behaviour. As GNU/Linux is a successful operating system, the way of the GNU to implement commands is something similar to a standard. Customer is the king and so we included a large amount of freeware tools into Solaris .

But where are the GNU tools now? They are seperated from other commands (same concepts as with the other toolsets). You will find them under /usr/sfw. Under this directory you will find freeware tools provided by Sun. In the case you wand a GNU make you will find it here. I will not accept any complaints about missing GNU tools when you didn't looked in this directory before. Important to know: We renamed some of the tools to seperate them from other versions (e.g. make and gmake)

You just have to include /usr/sfw in your PATH. The reason for not making this the default? The binary compatibility guarantee again. The default configuration of the operating system must enable the OS to be binary compatible to older versions of Solaris.

I hope this article shed some light into this topic

Posted by Joerg Moellenkamp in General at 21:00

Friday, October 10. 2008

## **Over 900 subscribers**

c0t0d0s0.org just reached the number of 900 subscribers:Really cool ...

Posted by Joerg Moellenkamp in About this blog at 21:10

## **Separated ZIL on ramdisk.**

I used ramdisks today in a live presentation of ZFS, L2ARC and sZIL (you can't use files for sZIL and L2ARC testing, you need a device). One of the customers asked me if there is a real use case for this configuration. My answer was: Not really. But this question haunted me the whole day while sitting in the train.

I had some weird ideas like using an UPS backed server to act as a sZIL device by providing access to a ramdisk via Fibre Channel (by using the FC target in COMSTAR) , but i had no idea if this would be really clever idea. Ramdisk based sZIL? Okay, i know there are now several admins with a cardiac arrest out there. The problem of a ramdisk based sZIL is the volatile nature of the ramdisk. When you loose the power on your systems, you have still a consistent pool but you loose the transactions stored in the sZIL.This is a less than desirable effect. Between you and loosing transactions is just the power switch There was a missing piece...

The good thing about a large community: Sometimes someone had already similar ideas long ago and had some time to do some experiments. So i found a really neat idea on the zfs-discuss mailing list with the important extra thought compared to my "dozing in the train"-idea.

The idea of Chris Greer has this nice additional twist making the concept of using a ramdisk sZIL more sensible. He writes in "An slog experiment (my NAS can beat up your NAS)": So I tried this experiment this week...

On each host (OpenSolaris 2008.05), I created an 8GB ramdisk with ramdiskadm. I shared this ramdisk on each host via the iscsi target and initiator over a 1GB crossconnect cable (jumbo frames enabled). I added these as mirrored slog devices in a zpool.

This is really a neat trick. You can mirror a sZIL device so you could distribute the ramdisks over several systems. When one of the system fails, you have still the other device with the same data. This configuration had an impressive effect on the performanceThe big thing here is I ended up getting a MASSIVE boost in performance even with the overhead of the 1GB link, and iSCSI. The iorate test I was using went from 3073 IOPS on 90% sequential writes to 23953 IOPS with the RAM slog added. The service time was also significantly better than the physical disk.I think you can drive this concept even farther by substituting the iSCSI over TCP over Gigabit with iSER (iSCSI over RDMA over Infiniband) and using separate small x86 systems for this task, each backed by a small UPS keeping the system up until an dd from the ramdisk to a hard disk has completed.

Chris, sounds like a really clever idea ...

Posted by Joerg Moellenkamp in Solaris at 20:28

## **Spleen**

I was in the mood for back-biting myself for the fact, that i tried to save some money for Sun. This meant: As there was no economy class seat in the plane from Dresde to Hamburg i had the opportunity to choose between Business Class (1h flight time) for hefty 400 Euros or the train in the second class for 48 Euros (5:50 ride time)

After not really thinking a lot about it, i've choosed the later option. But after holding a presentation i thought a little bit different about sitting five hours in a train.(I'm not really sure if this is cheaper, as i'm at home at 20:00 pm instead of 17:30pm, but that's a different question)ö But one fact made this a little bit better for me: Leipzig has a Starbucks now. It's in a few minutes walking distance to the main station. So i was able to indulge my spleen: I was able gather my 16th Starbucks mug now

Posted by Joerg Moellenkamp in Business Travel at 20:06

Thursday, October 9, 2008

### **MSI Wind PC and Opensolaris**

I just bought a MSI Wind PC . This is a device in the so-called "Nettop" range. 1.6 GHz Atom, 1 GB RAM, 320 GB harddisk and 30 Watts of power consumption in a small case. I bought it as the foundation of my new home storage server. I just put a first Opensolaris test installation on the system. I did a test installation of Opensolaris Build 99. It detected the network card, i was able to ping my router, the GNOME desktop was configured with the correct resolution (1440\*900). I was really pleased by this experience, as the RealTek 8111C NIC was known for making problems with older releases of Opensolaris.

Over the next days day i will transform this into a storage server with ZFS, SamFS et al. The internal disks will be the cache for SamFS, the external USB disks will be the archive storage for this system. I hope to put the archive harddisks to sleep at most of the time by this configuration due to activated power management.

Posted by Joerg Moellenkamp in Solaris at 21:53

Wednesday, October 8, 2008

## Anmerkungen zum ZFS-Tutorial in der c't - revisited

Eigentlich wollte ich das Thema "ZFS-Tutorial in der c't" mit dem letzte Artikel abschliessen, aber meine Anmerkungen haben doch so viele Reaktionen zur Folge gehabt, das ich mich dazu entschlossen habe, mit einem Artikel darauf zu reagieren, um noch mal einige Dinge zu erlaeuern.

Es scheint da ein Missverstaendnis zu geben. Das von der c't beschriebene Verhalten tritt nicht schon dann auf, wenn einfach eine Platte ausfaellt. Die man page zu zpool schreibt dazu bei der beschreibung zu failmode:Controls the system behavior in the event of catastrophic pool failure.Die Betonung liegt hier auf catastrophic pool failure. Dies ist immer gegeben, wenn nicht mehr ausreichend Redundanzen zur Verfuegung stehen, um den Pool aufrecht zu erhalten: Bau ich ein RAID ueber 3 Platten, und ziehe eine Festplatte, dann habe ich genuegend Redundanzen um den Betrieb weiter sicherzustellen. Ziehe ich zwei Festplatten, dann habe ich nur noch eine Festplatte, der Pool ist nicht mehr betriebsfaehig, der im letzten Text beschriebene failmode wird wirksam. Vor Build 77 und darauf basierender Distributionen haette uebrigens an der Stelle das System stumpf gepanict. Das war vorher die Defaulteinstellung. Wie ich schon mal schrieb, Panics sind nicht dazu einfach nur Ungemach zu verbreiten, sie sind dazu entwickelt worden, den Zustand der Daten auf der Platte zu schuetzen.

Wann passiert sowas? Wenn ich beispielweise der Meinung bin, das alle meine Spiegel hinter dem selben Kabel am selben Controller stecken muessen. Oder vielleicht viele Controller habe, aber durch eine wenig zielfuehrende Konfiguration beide Spiegelhaelften hinter einem Controller habe. Macht der Controller oder das Kabel Probleme, sind auf einem Schlag weniger Platten da, als zur Betriebsfaehigkeit eines RAID noetig sind und zum Schutz der Daten geht der betroffene ZFS-Pool dann in den von der Option failmode angegebenen Zustand.

Das ist uebrigens auch der Grund, warum es eine ausnehmend schlechte Idee ist, aus einem RAID controller eine LUN rauszumappen, und da einfach nur ZFS drauf zuwerfen. Da kann ich gleich UFS auf das device kippen wenn man so einfach einige der wesentliche Vorteile von ZFS aufgibt (Checksummen koennen dann zwar feststellen, das etwas nicht stimmt, es aber mangels Redundanzen nicht korrigieren. Zudem kommt da genau obiges Problem bei rum. Die ganzen Platten womoeglich noch an einem Ghetto-RAID-Controller ohne Cache-Mirroring. Prelude to disaster.

Ich habe das Verhalten mal mit mehreren Files als vdevs nachgestellt (... habe leider mein Reise-Multipack im anderen Koffer ...). Es ist hier anzumerken ,das auch wenn der erste zpool status blockt ein in einem anderen Fenster durchgefuehrter zpool statusdurchaus schluessige Daten liefert:# zpool status

```
pool: test
state: FAULTED
status: One or more devices are faulted in response to IO failures.
action: Make sure the affected devices are connected, then run 'zpool clear'.
see: http://www.sun.com/msg/ZFS-8000-HC
scrub: scrub completed after 0h0m with 0 errors on Wed Oct 8 19:06:21 2008
config:
```

NAME	STATE	READ	WRITE	CKSUM
test	FAULTED	0	0	0 experienced I/O failures
raidz1	UNAVAIL	0	0	0 insufficient replicas
/root/test1	UNAVAIL	0	0	0 cannot open
/root/test2	UNAVAIL	0	0	0 cannot open
/root/test3	ONLINE	0	0	0

Alternativ ist hier auch ein Block in die Logfiles der Fault Management Architecture hilfreich:# fmdump -Ve -c "\*vdev.open\_failed\*"

```
TIME CLASS
Oct 08 2008 19:16:34.538548765 ereport.fs.zfs.vdev.open_failed
nvlst version: 0
class = ereport.fs.zfs.vdev.open_failed
ena = 0x3670c5e7f4b00001
detector = (embedded nvlst)
nvlst version: 0
version = 0x0
```

```
scheme = zfs
pool = 0x950cc5a239b45bd7
vdev = 0xd17f1aa49e4ae02b
(end detector)
```

```
pool = test
pool_guid = 0x950cc5a239b45bd7
pool_context = 0
pool_failmode = wait
vdev_guid = 0xd17f1aa49e4ae02b
vdev_type = file
vdev_path = /root/test1
parent_guid = 0x27919fd695c7d103
parent_type = raidz
prev_state = 0x1
__ttl = 0x1
__tod = 0x48ecef2 0x20199a1d
```

Oct 08 2008 19:17:27.703175938 ereport.fs.zfs.vdev.open\_failed

nvlist version: 0

```
class = ereport.fs.zfs.vdev.open_failed
ena = 0x3706742200c00001
detector = (embedded nvlist)
nvlist version: 0
  version = 0x0
  scheme = zfs
  pool = 0x950cc5a239b45bd7
  vdev = 0x6e1d53e952637ff4
(end detector)
```

```
pool = test
pool_guid = 0x950cc5a239b45bd7
pool_context = 0
pool_failmode = wait
vdev_guid = 0x6e1d53e952637ff4
vdev_type = file
vdev_path = /root/test3
parent_guid = 0x27919fd695c7d103
parent_type = raidz
prev_state = 0x1
__ttl = 0x1
__tod = 0x48eceb27 0x29e99d02
```

#Wie bekomme ich das nun wieder in Betrieb, ohne zu rebooten? Platten wieder anstecken und der Befehl zpool clear ist danach dein Freund. Steht ebenfalls in der manpage zu failmode=wait.

Posted by Joerg Moellenkamp in Solaris at 19:36

## links for 2008-10-08

[Exposing a MySQL Database with RESTful Web Services - NetBeans IDE 6.0 Tutorial](#)

(tags: webservices webservice tutorial restful rest netbeans)

[Getting Started with RESTful Web Services on Glass Fish - NetBeans IDE 6.0 Tutorial](#)

(tags: webservices tutorial rest programming restful netbeans)

[radiocheck - Funksprache wie sie vielleicht real gesendet wurden](#)

(tags: aviation)

AMD und ATIC grÄ¼nden "The Foundry Company" - Planet 3DNow! - Das Online-Magazin fÄ¼r den  
AMD-User

(tags: amd foundry)

Posted by del.icio.us at 13:00

Tuesday, October 7, 2008

## **Anmerkungen zum ZFS-Tutorial in der c't**

Die c't schreibt im Artikel ueber ZFS: Aber auch ein RAID-Z kann richtig kaputtgehen: Als wir im laufenden Betrieb aus einem RAID-Z1-Verbund mit drei Platten zwei entfernten und mit `zpool status` den Status des Speicherpools erfragen wollten, hing der `zpool` befehl bei maximaler Plattenaktivität - und das stundenlang. Auch andere Befehle wie `df` oder `fs`, die Informationen des Dateisystem erfragen, blieben hängen. Selbst ein Shutdown gelang nicht mehr, erst nach einem harten Reset ließ sich der defekte Speicherpool retten. Okay, dieses Verhalten von ZFS ist ein erwuenshtes Verhalten. Das RAID-Z ist in diesem Moment nicht mal kaputt. Nein ... das ist kein Schönreden eines ZFS fanboys. Ich will das mal im folgenden Text erlaeuern.

Man kann ein Filesystem in zwei Richtungen designen. Das eine Designdogma ist "Datenkorrektheit um jeden Preis", das andere ist "Datenverfügbarkeit um jeden Preis". Man kann nicht beide gleichzeitig erfüllen. Wenn ich "Datenkorrektheit" als oberstes Ziel setze, so muss ich gegebenenfalls den Zugriff auf die Daten unterbinden. Wenn ich die Datenverfuegbarkeit an oberste Stelle stelle, muss ich unter Umständen auch einmal Daten liefern, die von einem Array mit unklarem Zustand kommen.

ZFS hat sich zum Ziel gesetzt, die Datenvalidität als oberstes Kriterium zu sehen. Das hat ganz praktische Gründe: Datenvalidität sicherzustellen heisst auch, das ich mir gegebenenfalls den Filesystem check auf einem mehrere Petabyte grossen Filesystem sparen kann. Viele Entscheidungen in ZFS sind in Hinblick auf die unschönen Effekte ausgelegt, die 200 TB Filesysteme so mit sich bringen koennen.

Um mal kurz abzuschweifen: Verabschiedet euch von der These, das Daten auf rotierendem Rost sicher sind. Ich finde es teilweise höchst erstaunlich, wie gross das Vertrauen in eine Technik ist, die auf unterster Ebene ein Technik einsetzt, die in ihrer Bezeichnung das Wort "Wahrscheinlichkeit" beinhaltet. Bei der Beschreibung von PRML wird einem doch warm ums Herz. Der weg von der magnetisierten Blase Eisenoxid bis hin in den Speicher ist lang und tierisch gefährlich für ein Bit. Verabschiedet euch insbesondere, das Daten sicher sind, wenn ihr sie auf consumer-grade Elektronik speichert. Zig Komponenten und alle vom billigsten Hersteller. Gerade soviel Standard und Testen, als das man nicht sofort damit beim Kunden auf die Nase faellt. IT ist oftmals heute nur noch best-effort computing ... soviel zum allgemeinen IT-Rant ... zurueck zum Thema. ZFS wurde in Hinblick auf diese Entwicklung hin entworfen .

Der Autor des Textes in der c't hat also mal ein paar Platten gezogen. Das setzt voraus, das man das auch kann und darf. Das geregelte Entfernen von Festplatten kann durchaus etwas problematisch sein : IDE unterstützt beispielsweise Hotplug ueberhaupt nicht (Hotplug heisst ja auch Hotunplug). Damit das funktioniert, muessen eine ganze Reihe von Dinge erfüllt werden. Der Controller muss im AHCI-Modus laufen. Laeuft beispielsweise ein SATA-Port im "IDE emulation mode" kann man zu diesem Zeitpunkt beispielsweise Hotplug vergessen. Die Verbinder mögen vielleicht elektrisch hergeben, das man sie während des Betriebes zieht. Aber die Meldung des Controllers, das da gerade etwas passiert ist, findet nicht statt. Besonders bei SATA gibt es viele Fallen, in die man laufen kann.

Das ist uebrigens einer der Gründe warum SAS selbst bei einfachen Ghetto-JBODS verwendet wird (eSATA ist Desktopstorage ... don't use it out of home). Es ist dort einfach sauberer und besser standardisiert und implementiert. Damit Hotplug und -unplug wirklich funktioniert, muss das auf der ganzen Kette von der Festplatte bis hin zum Treiber sauber implementiert werden. Es steckt viel Testerei in der Realisierung Storage-Systemen und je consumer-grade die Technik ist, deso mehr testen muss man reinstecken, das das ordentlich laeuft. Sind die Standards eingehalten, funktioniert ein Modus fehlerfrei ... uswuf.

Und das ist teilweise eben nicht der Fall, insbesondere wenn man teilweise einfach nur die Ports in einem Mainboard verwendet. Manchmal liegt es sogar einfach nur an den BIOS-Einstellungen, das dann irgendwas sich anders verhaelt als erwartet (IDE emulation anstatt AHCI). Oder manchmal muss man sogar zwangsweise auf IDE-Emulation zwangsweise umschalten, weil ansonsten der Chip nicht sauber funktioniert. Man kann also mit solchen Problem sehr einfach sein IDE-Subsystem in einen Zustand bringen, bei dem man beispielsweise trotz Hotplug die Finger von den Kabeln lassen muss.

Okay ... zurueck zum stehenden `zfs status`. Was macht ZFS jetzt per default, wenn es Ärger im IDE-Subsystem gibt? Es geht davon aus, das jeder weitere Zugriff auf das System potentiell den Inhalt des Filesystem schädigen kann und unterbindet daher jeden Zugriff darauf, bis entweder die fehlenden Devices wieder anschlossen worden sind beziehungsweise diese ausgetauscht worden sind (z.B. automatisch durch die Fault Management Architecture)

Jetzt werden einige Leute sagen, das ihnen die Datenkorrektheit egal ist und auf alle Faelle auch bei unschoenen Betriebszuständen weiter Daten zumindestens gelesen werden sollen. Nun, man kann dieses Verhalten ändern. Dazu gibt es das failmode property eines Storage pools. In der man page zu zpool steht zum Thema des failmode:failmode=wait | continue | panic

Controls the system behavior in the event of catastrophic pool failure. This condition is typically a result of a loss of connectivity to the underlying storage device(s) or a failure of all devices within the pool. The behavior of such an event is determined as follows:

waitBlocks all I/O access until the device connectivity is recovered and the errors are cleared. This is the default behavior.

continuereturns EIO to any new write I/O requests but allows reads to any of the remaining healthy devices. Any write requests that have yet to be committed to disk would be blocked.

panicPrints out a message to the console and generates a system crash dump.

Warum funktionieren aber nun andere Filesysteme ohne solche Spirenzchen. Es ist das allgemeine Designdogma was hier anders ist. ZFS misstraut der Hardware und die Defaulteinstellungen sind Zeichen dieses Misstrauens. Im Grunde misstraut es auch den Treibern in Solaris. Um die Datenvalidität zu sichern, geht ZFS soweit den Zugriff auf die Daten durch Blocking zu verweigern, bis das System wieder in einem definierten Zustand ist, von dem ausgehend man weiterarbeiten kann.

Linux, ext2/ext3, NTFS, Windows gehen von einem anderen Modell aus, reagieren anders, solange da etwas ankommt, wird ausgeliefert. Ob das allerdings bei einem Filesystem, das nicht ueber Checksummen verfügt, um die Datenvalidität sicherzustellen, eine wirklich kluge Idee ist, mag jeder für sich selbst entscheiden.

Kabelziehen ist sowieso noch ein recht einfacher Fehler. Die interessante Frage ist doch, ob man einem Festplattensubsystem beim Lesen vertrauen kann, bei dem gerade mal eine Platte ohne Mitteilung verschwunden ist. Die Entscheidung, dieses Risiko einzugehen, sollte eine bewusste Entscheidung sein, ergo die oben gewählte Defaulteinstellung.

Achja ... ein Wort noch zu den Tests, die Festplattenausfälle durch das Ziehen von Kabeln simulieren wollen: Ihr testet damit abfallende Kabel und den Effekt von Benutzern, die wild einfach mal so Kabel ziehen, aber keine defekte Festplatte. Wer allerdings ein Problem mit abfallenden Kabeln hat, sollte sich ueberlegen, ob er nicht mal anständige Kabel einsetzt, so mit Verschraubungen oder vernuenftigen Arretierungen.

Posted by Joerg Moellenkamp in Solaris at 21:08

### **TheRegister about the NTAP/Sun lawsuit**

The Register writes in NetApp faces Sun lawsuit loss:NetApp was unable to comment immediately on this story. Sun's win - if it is a win - and the PTO decisions potentially turn NetApp's WAFL IP into, well, IP waffle. Sun gets a welcome PR boost to its ZFS and open source credentials, leaving NetApp with a bloody nose and a 22-patent IT infringement case to deal with. Oh, and the economy is going down the tubes too. It isn't so sunny in Sunnyvale right now.

Posted by Joerg Moellenkamp in The IT Business at 14:32

### **links for 2008-10-07**

767 deal structures studied for Boeing's ANA/JAL 787 compensation deals

(tags: 767 787 Boeing Delayliner)

Posted by del.icio.us in del.icio.us at 13:00

### **ZFS in der c't**

In der aktuellen Ausgabe der c't (21/2008 vom 29.9.2008) kann der geneigte Leser auf den Seiten 194 bis 199 ein gutes Tutorial zum Thema ZFS finden. Positiv geschrieben. Leider hat die c't eine Chance ausgelassen. Zwar ist in der Zeitschrift auch ein Bericht auch ueber Solid State Disks, aber im ZFS-Bericht fehlt der Hinweis auf die Konfiguration von sZIL und L2ARC (zugegebenermassen erwaeht der Autor es im letzten Absatz). Hätte vielleicht auch zu grosse

Begeisterungstuerme in der Leserschaft pro ZFS ausloesen koennen

Posted by Joerg Moellenkamp in Solaris at 10:23

Monday, October 6. 2008

### **Earworm**

Sh... can't get this tune out of my head ...

Posted by Joerg Moellenkamp in Music at 17:41

### **links for 2008-10-06**

Finanzkrise in Deutschland - Demokratiealarm - Wirtschaft - sueddeutsche.de

Muss ein Schaediger nur dreist genug und der Schaden nur gross genug sein, damit der Staat das Desaster nobilitiert?

(tags: wirtschaft prantl kreditkrise)

Posted by del.icio.us in del.icio.us at 13:00

### **Oracle on CMT**

Glenn.Fawcett of Sun and Andrew.Holdsworth of Oracle held a presentation about the performance characteristics of Oracle on CMT: Growing Green Databases with Oracle and Sun UltraSPARC T-series servers. The presentation offers an interesting insight: The higher the amount of parallel session the more favorable the number of yielded transactions per seconds. Additionally the presentation offers some handy tuning tips for running Oracle efficently on CMT systems.

Posted by Joerg Moellenkamp in Sun at 09:58

Sunday, October 5. 2008

### **News from the NetApp vs. Sun lawsuit**

Mike Dillon published some news in his blog about the NTAP vs. Sun. The development looks really positive for Sun: Most significantly, the Court found each of the asserted claims in NetApp's 7,200,715 patent relating to RAID technology to be "indefinite" - [...]. With regard to NetApp's '715 patent, the court agreed with Sun's position that the claims of the patent are flatly inconsistent with and impossible under the teaching of the patent specification. In effect, unless NetApp appeals and this finding is reversed, the '715 patent is effectively invalidated in this case and against others in the future. and In addition, the Court's findings on the terms "server identification data", "domain name", "portion of a communication" "element of a communication" and "completing a write operation within a local processing node" further strengthen our position that the processors, network interface and systems management software used across NetApp's product line infringe Sun's patents. The ZFS information page is already updated in regard of this new documents.

Posted by Joerg Moellenkamp in Sun at 23:25

### **Heise und der SAP SD Benchmark**

Da muss ich doch meinen Hut vor Herrn Stiller ziehen. Ich hatte mich ja ein wenig ueber die Bewertung der unterschiedlichen SAP-SD Benchmarks fuer die Hexacore Systeme echauffiert. Dies hat Herr Stiller nunmehr in der neuesten Ausgabe des Prozessorgefluesters korrigiert: Konfusion, unter anderem in meinem Hirn, erzeugten SAP-SD-Benchmarkergebnisse, die Sun für Systeme mit Intels neuen Hexa-Core-Xeons eingereicht hatte, denn sie lagen weit unter denen, die Konkurrent HP für seine ProLiant-Server nannte – bei gleichem Prozessor und Chipsatz. Der Unterschied rührt dabei weder von der Hardware noch von den unterschiedlichen Betriebssystemen und Datenbanken her (Solaris 10 mit MaxDB7.6 hier, Windows Server 2003 und MS SQL Server 2005 da), sondern im Wesentlichen vom Einsatz von Unicode (UTF-16) bei SAP-ERP und in der Datenbank. Bisläng hat einzig Sun Ergebnisse mit Unicode-Systemen eingereicht, die laut SAP eine bis zu 30 Prozent höhere CPU-Leistung erfordern. Die Ergebnisse sind somit nicht mit anderen vergleichbar. Vielen Dank dafür an Herrn Stiller

Man kann sich bei der Angelegenheit aber durchaus fragen, warum ausser Sun alle Hersteller den in der Praxis nahezu irrelevanten Wert fuer non-unicode SAP in Benchmarks zertifizieren lassen. Man kann vielleicht sagen, das das Verhalten von Sun marketingtechnisch unschoen ist, aber wir sind eben eine "engineer company" und damit teilweise einfach viel zu ehrlich ...

Posted by Joerg Moellenkamp in Sun at 16:03

### **links for 2008-10-05**

digital:pardoe â€“ iSynct

(tags: sync software productivity osx nokia macosx)

Posted by del.icio.us in del.icio.us at 13:00

Saturday, October 4, 2008

## **Going to lunch - with the eyes of an IT architect**

When you work as an architect in the IT industry you tend to look at the non-IT processes in your world with the eyes of an IT architect. And at many occasions you find the same problems. Perhaps the design techniques of efficient CPUs and software can help to do a really efficient lunch distribution ... and i'm sure some of the problems of CPU or software design were solved by observing the people at lunch time.

Let's have just a look at the lunch shops around the Sun office in Hamburg.

**Mickleys** - Problems with singletonsMickleys is a Turkish deli. This is a nice example for problems introduced by singletons. The processing ("giving you the food") is really efficient. But the cashier is implemented as a singleton, just one cashier but up to three processing threads. The processing threads idle most of the time because of the cashier thread blocking the queue. After a short time the queue gets so long that the processing threads can't dispense any data ... eer ... food.

**Sultans** - Stateless session transferSultans is a doner shop. They have the same singleton problem like Mickleys but the cashier process is really efficient. But Sultans is a good example for a stateless transfer of sessions between two services. You can look at Sultans as a service separated into two subservices: The cashier service and the doner making service. When you pay for your doner you open the session, but the data of the "session context" isn't distributed to "doner making service". After paying for your doner your session will be requeued to the "doner making service" queue and you specify your doner structure again ("doner plate with fries and spicy sauce"). By doing so the waiting requests can be easily loadbalanced to several doner making services (the one at the Sultans near the office has two doner making services executing independently from each other). By the way: The doner plate at Sultans is really a good lunch.

**Oh it's fresh** - Scheduling in multiprocessor systemsOkay ... at "Oh it's fresh" (OIF in the following text, but it's called HAM03 sometimes, over the day you find at least one Sun employee there making a short break at any time) you find a good example of process scheduling in multiprocessor systems. At OIF you have two execution units consisting of a CPU (cash processing unit) and a FPU (food processing unit) each. You have a large cache with preprocessed food. All consumer processes are queued in a run queue. All consumers are in a state where they are runnable but want to be scheduled to cash processing unit. When both cash processing units are functional, there is a run queue for every cash processing unit. Often the consumers can be served without waiting for other high latency. As soon as the consumer is scheduled to a Cash processing unit. As long as the lunch is in the cache, the process is completely executed without leaving the cash processing unit. Sometimes there is a high latency event. Sometimes there is a high latency event. For example, when you order a Latte, the job is transferred to the "Coffeemaking Offloading Engine", the consumer process gets preempted and the process is put on the sleep queue. Now the next consumer process is scheduled to the CPU. As soon as the high latency event is over (the Coffeemaking Offloading Engine has prepared the Latte), the actual consumer is rescheduled to the run queue with a priority as high as the actual process. Directly after the completion or the transfer of the new customer to the sleep queue, the longest waiting consumer waiting for the deliverable resource is executed again. It's like in Solaris ... there are conditions in the sleep queue. Only the matching consumers are woken up by seeing a glass of Latte, nevertheless the scheduler is aware of the priority of the consumer in the sleep queue.

I don't really like this shop. Some of the people over there got overconfident about the fact, that they work at a relatively hip deli there. I don't like to get insulted by a guy with a third of my IQ and a tenth of my annual income. When you get violent thoughts like jumping over the counter and putting the mug of coffee into a body hole that wasn't invented for consuming coffee, you should choose a different coffee-dealer shop. By the way: The coffee is horrible over there ... just a wad of water with homeopathic doses of coffee beans used in the process. The substance over there isn't far away from the sign "No coffee beans were harmed during cooking this coffee". Just walk hundred meters and you got much better coffee and a much friendlier CPU/FPU.

**Pic-a-deli** - The thundering herd problemOIF has a well implemented lunch scheduling mechanism. Now I want to get to an inefficient lunch scheduling. This scheduling doesn't now priorities. So you have an interesting effect in the design of this scheduling mechanism. It's called "thundering herd". The Pic-a-deli has this problem (BTW: all scheduling mechanisms are affected by the "thundering herd" problem, the question is just the severity of the effect). At first you are in the run-queue for the cash processing unit. After this the guests are moved to sleep-queue. The food processing unit is somewhat chaotic. The FPU yells "Pasta", all consumers who have ordered "Pasta" on the sleep queue wake up

and fight for the access to the "Pasta". The scheduling of the FPU is unaware of the sequence thus processes can really starve in the sleep queue. This problem is really similar to the "thundering herd problem". It's not uncommon that a colleague gets it food when others are already finished, despite of ordering it at almost the same time ...

Mc Donalds - Speculative Execution/Branch Prediction  
When you eat at Mc Donalds you see the effect of branch prediction. Let's assume you want to order a meal ... let's assume you want to order a BigMac (a quarter-pounder with cheese) ... most of the times this is a low-latency event, you get your fries and your burger within seconds, but when you order an McRib you wait a few minutes. This is branch prediction and speculative execution in action. The staff at speculatively executes the preparing of BigMacs, so the cache (the hot bay behind the desk) is filled with a certain amount of BigMacs at any time. In 90% of all cases the consumer decide for a Big Mac. Thus ordering a BigMac is a low-latency event. But in 10% the consumer threads orders a McRib. Well ... now you you have high-latency event ... the cache doesn't contain any McRibs. When there is nothing much to do, you stay at the top of the queue, but the pipeline is stalled for this moment. At lunch time, the Cash Processing Unit does a context switch and processes the next consumer process. When the high-latency event of preparing a McRib is completed, the Cash Processing Unit switches back to the waiting process after the new process has completed. The new process isn't preempted by the completion of the high-latency event.

McDonalds - Scout Threading  
McDonalds is good example for Scout Threading, too. Let's assume you are stuck at the telephone, but a group of colleagues are almost starved, thus they can't wait any longer, but they are willing to order a McRib for you. You leave the office two minutes later. As you arrive at McDonalds the cache is already warmed, as the scout colleagues already ordered a load of a McRib into the cache two minutes earlier. The complete high-latency event of preparing a McRib is hidden by the scout colleagues doing the upfront order.

The concept of scout threads in computing is exactly the same. A scout thread runs 100 cycles or so upfront to the real thread and warms the cache, so the real thread doesn't have to wait for the data.

Posted by Joerg Moellenkamp in Computing at 15:53

Friday, October 3. 2008

### **IBM stock under pressure**

In the last few days the stock of IBM was hit quite hard. Two days with roundabout 5 percent decrease. You can wonder a little bit about this, but there is an interesting article in the "Between the lines" blog: What's really ailing Big Blue shares? Hint: IBM is part bank. Larry Dignan writes: Big Blue's financing unit, which leases hardware and finances projects, is big enough that the Securities and Exchange Commission put IBM on the "do not short" list. This list is designed to get shorts-folks that bet against stocks-off the backs of financial services companies long enough to raise capital or at least survive. Thus IBM is possibly hit by two trucks: By economic slowdown as a computer ... sorry ... consulting company and by the credit crisis as a financial company. The article is really an interesting read.

Posted by Joerg Moellenkamp in The IT Business at 13:31

Thursday, October 2, 2008

## links for 2008-10-02

Self-service, Prorated Super Computing Fun! - Open - Code - New York Times Blog

(tags: webservices scalability storage s3 python programming performance parallel mapreduce hadoop grid java)

Introduction to Hadoop

(tags: mapreduce hadoop storage grid java)

Meet Hadoop - Part 2

(tags: mapreduce hadoop storage grid java)

Meet Hadoop - Part 1

(tags: mapreduce hadoop storage grid java)

Celeste at OpenSolaris.org

(tags: sun storage software distributed data backup opensource)

Welcome to Hadoop!

(tags: mapreduce hadoop storage grid java software programming performance server opensource)

Concept to Reality: Deep-Stall Avoidance

(tags: stall aviation)

Asymmetric Collateral Damage

Basel II, the Mortgage House of Cards, and the Coming Economic Crisis

(tags: crisis finance)

Posted by del.icio.us at 13:00

## Digging into Apache Hadoop

I'm exploring Apache Hadoop at the moment. This looks like a really interesting technology. What's Hadoop? Hmm ... to explain it in a really simplified manner: It's a distributed, highly available datastore. Okay ... yawn ... no big deal so far.

But there is an interesting twist in Hadoop. Let's assume you have vast amounts of log files. A pile of data in the size of multiple Terabytes. You want to know the URLs of the Top-10 pages of your website.

The standard old-school approach to this problem is: Starting an analyser on a big server which gathers all data via block or file based protocols to this analysing server. Of course this approach has several bottlenecks: The size of the network pipes, the amount of computing power in a single box, the amount of IOPS in a single server, the amount of IOPS of a single storage attached to the server, the amount of memory in a single server.

But now think in HPC terms about this problem: You could divide this task in several ones. Let's assume 64 MB shards. You could compute the result for each of the shards on a separate node. To stay in our example: This step print outs the pageviews of any URL in it's shard. This fragments of the final result are collected and reduced to the final result: For example by adding the pageviews of a URL from every shard. By using the concepts you separate the task of

analysing the log files on thousands of nodes in parallel. You get your answer in minutes, not hours or days. The advantage of doing so is to bring the data intensive parts of computation to the data instead of bringing the data to the computation.

Such algorithms are called MapReduce. This concept was introduced by Google and the core competency of Google is to analyse big piles of data, thus such a mechanism is quite handy. I have several usecases in mind for such a solution: Commercial data warehousing, billing of large heaps of Call Data Records, mass converting jobs ... and so on ...

What has all this stuff to do with Hadoop. Hadoop is an open-source implementation of this concepts. The Hadoop Wiki writes: Hadoop is a framework for running applications on large clusters built of commodity hardware. The Hadoop framework transparently provides applications both reliability and data motion. Hadoop implements a computational paradigm named Map/Reduce, where the application is divided into many small fragments of work, each of which may be executed or reexecuted on any node in the cluster. In addition, it provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both Map/Reduce and the distributed file system are designed so that node failures are automatically handled by the framework. It does the housekeeping, the separation of data in shards, the distribution of the analysing tasks on the server. You can view at it as an API/command line controlled grid engine for data distribution and data processing.

It consists out of the Hadoop Core, the Hadoop Distributed File System (it's not a POSIX filesystem integrated to the VFS, you can think of it like FTP, you need a client or you use an API to use it), there is even a scripting language helping you to write the analysing jobs. This language is called Pig. Additionally there is an effort to implement a database for structured data on top of Hadoop with HBase.

At the moment there are some gotchas in this technology. For example you can't work with compressed files in it, as gzip files shards aren't decompressible separately (okay, it's a problem of gzip, but it prevents you to work with it) (Update: I'm not entirely sure about this, it seems that you can work with block based compression like bzip2, and gzip is in development, at least according to the Pig documents). But here ZFS on-the-fly compression can be very helpful. I think I will create a Hadoop testbed with multiple zones on one of my systems this weekend.

Posted by Joerg Moellenkamp in Technology at 11:07

Wednesday, October 1. 2008

## **Open Storage Summit Keynote**

Ben Rockwood gave a really interesting keynote at the first Open Storage Summit about "Storage in the Cloud". When you are interested in storage with Opensolaris it's definitely worth a look.

Posted by Joerg Moellenkamp in Solaris at 19:43

## **Who needs enemies ... - Part 2**

Kollege 1: "... concrete actions ..."

Kollege 2: "Seit wann machen wir Beton-Aktionen? ;)"

Posted by Joerg Moellenkamp in Sun at 14:28

## **links for 2008-10-01**

Downgrading 2.0

(tags: iphone)

Posted by del.icio.us in del.icio.us at 13:00

## **Bull chosen for 200-Teraflops Supercomputer**

Good news for Sun(NASDAQ: JAVA) : The Forschungszentrum Julich in Germany Chooses Bull to Deliver a 200-Teraflops Supercomputer for the JuRoPa Project. JuRoPA stands for "Jülich Research on Petaflop Architectures".

Good news for Sun? Yes ... because Bull is mainly the prime contractor for this HPC project. Sun delivers large parts of the hardware and important components of the software: Julich's new supercomputer is a cluster combining Bull NovaScale servers, Sun blade servers, all based on Intel(R) Xeon(R) Nehalem processors, and complete HPC cluster software provided by Partec. This is combined with a high-performance input-output (I/O) system based on a Sun ZFS/Lustre file system, guaranteeing end-to-end data integrity. Based on the actual Top500 list, this would be the sixth largest supercomputer worldwide and the largest supercomputer in Europe.

Posted by Joerg Moellenkamp in Sun at 08:07