

# Hadoop

Jörg Möllenkamp Principal Field Technologist

**Sun Microsystems** 



#### Agenda

Introduction CMT+Hadoop Solaris+Hadoop Sun Grid Engine+Hadoop



# Introduction



#### l'm ...

Jörg Möllenkamp better known as "c0t0d0s0.org" Sun Employee Principal Field Technologist from Hamburg



#### l'm ...

Jörg Möllenkamp better known as "c0t0d0s0.org" Sun Employee Principal Field Technologist



thus a part of the HHOSUG as well ...



### An apologize right at the start ...



#### No live demonstration ...



# ....Sorry



# Had a "shortnotice" customer meeting at 10:00 o'clock ...

3 presos yesterday, one this morning. so my voice may be a single point of failure ...



#### Or to say it with Rudi Carrell "A moment ago in a meeting room in Bremen, now on the stage in Berlin"



#### Had no time to test my "demo case" ....



#### And i've learned a thing in thousand presos: Never ever do a live demo without tests ... ... will ruin your day big time ...

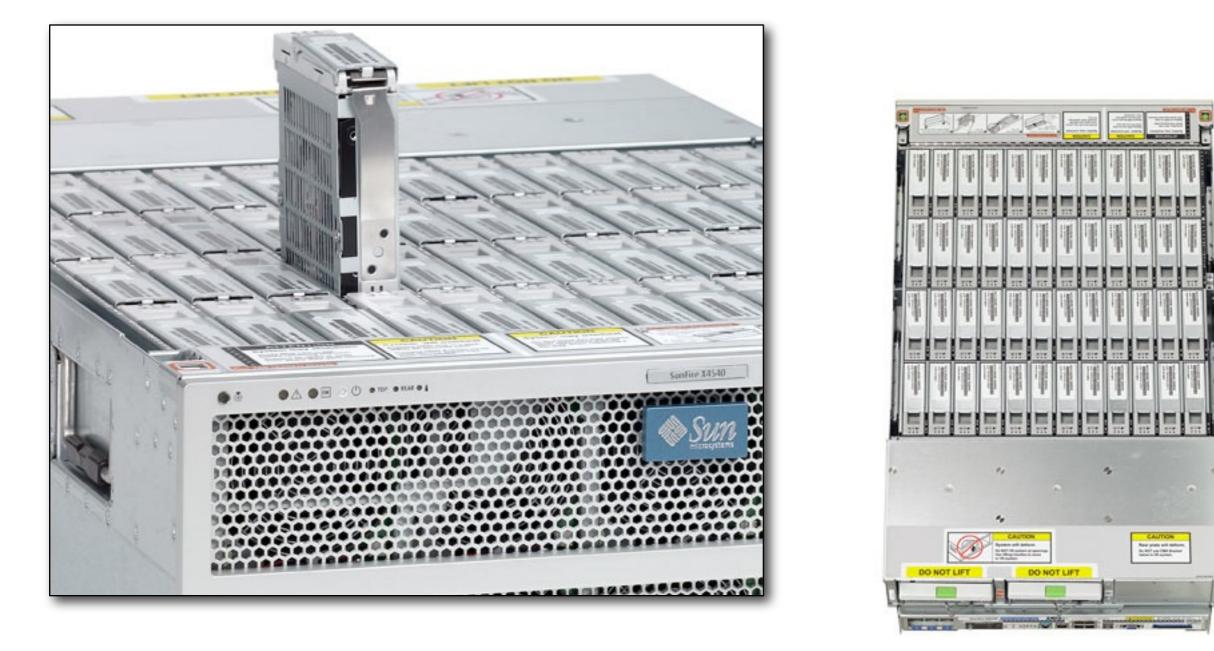


## In the scope of this presentation: Why is Sun interested in Hadoop? Mutual significance A little bit bragging about some new Sun HW



## Not in the scope of this presentation: Explaining you the idea behind Hadoop The History of Hadoop Just providing a list of Sun Hardware









## Sun+Hadoop



#### Why is Sun working with Hadoop?



### At first: It's an "I" technology.



### Not "I" for "Internet"



## "I" for "Interesting stuff"



### At the CEC2008 Hadoop was an important part on the Global Systems Engineering Tracl



#### We think that: Hadoop can provide something to Sun But as well: Sun can provide something to Hadoop



# Hadoop+CMT



## What can Hadoop provide for Sun?



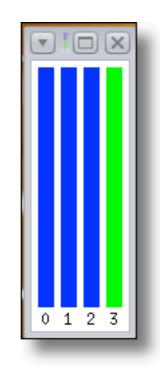
### Another usecase for a special kind of hardware



### CMT Chip Multi Threading



#### 4 or 8 Cores are for Sissys





#### 2005 UltraSPARC T1 8 Cores 4 Threads per Core 32 Threads per System



#### 2007 UltraSPARC T2 8 Cores 2 Integer Pipelines per Core 4 Threads per Pipeline 64 Threads per CPU



#### 2008

UltraSPARC T2+ CMT goes SMP 8 Cores 2 integer pipelines per core 4 threads per pipeline 64 Threads per CPU 4 CPUs per system 256 threads per system

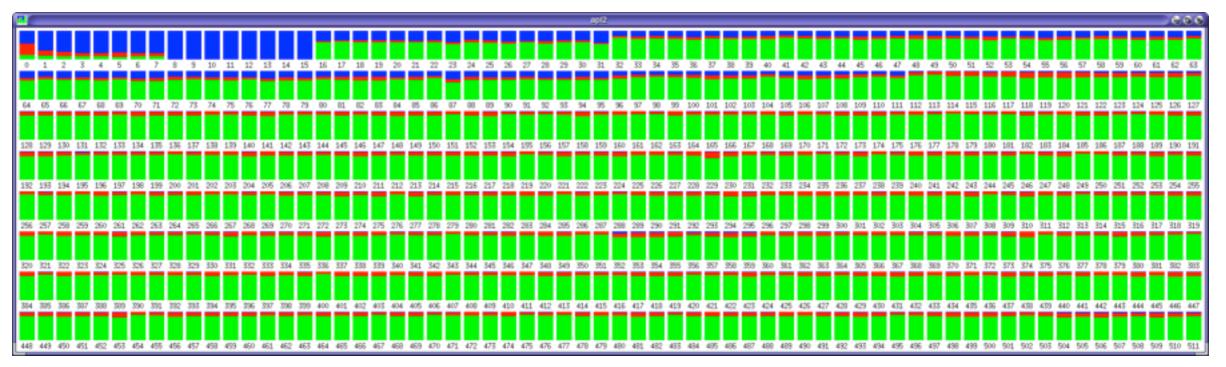


#### 2010

UltraSPARC "Rainbow Falls" 16 Cores 2 integer pipelines per core 4 threads per pipeline 128 Threads per CPU 4 CPUs per system 512 threads per system



#### That would look like that:





#### obviously a single grep process don't scale that well on this system ...



#### Those system eat threats ... lot's of them ...



#### Otherwise it's relatively slow ... just 1.6 GHz at the moment.



#### But 4 memory controllers today, more later ... because frequency means nothing if your proc has to wait for data from RAM ...



### Or perhaps a better analogy ... It doesn't matter if you stir your diner at 1.6 GHz or 4.7 GHz

when you have to wait for your significant other to get the bottle of wine from the cellar.





my colleagues made the last screenshot on this system



We have an operating system that can use this amount of threads.



### But that's only half of the story: You need applications that are able to generate the load.



## UltraSPARC Tx is a massively parallel, throughput centric architecture ...



### Sound familiar?



### Yes ... indeed!



### Would you like your Hadoop in a box?



### Wasn't Hadoop developed with small boxes in mind?



### Yes ... of course. But there is still a reason for using T-Class systems.



### Density!

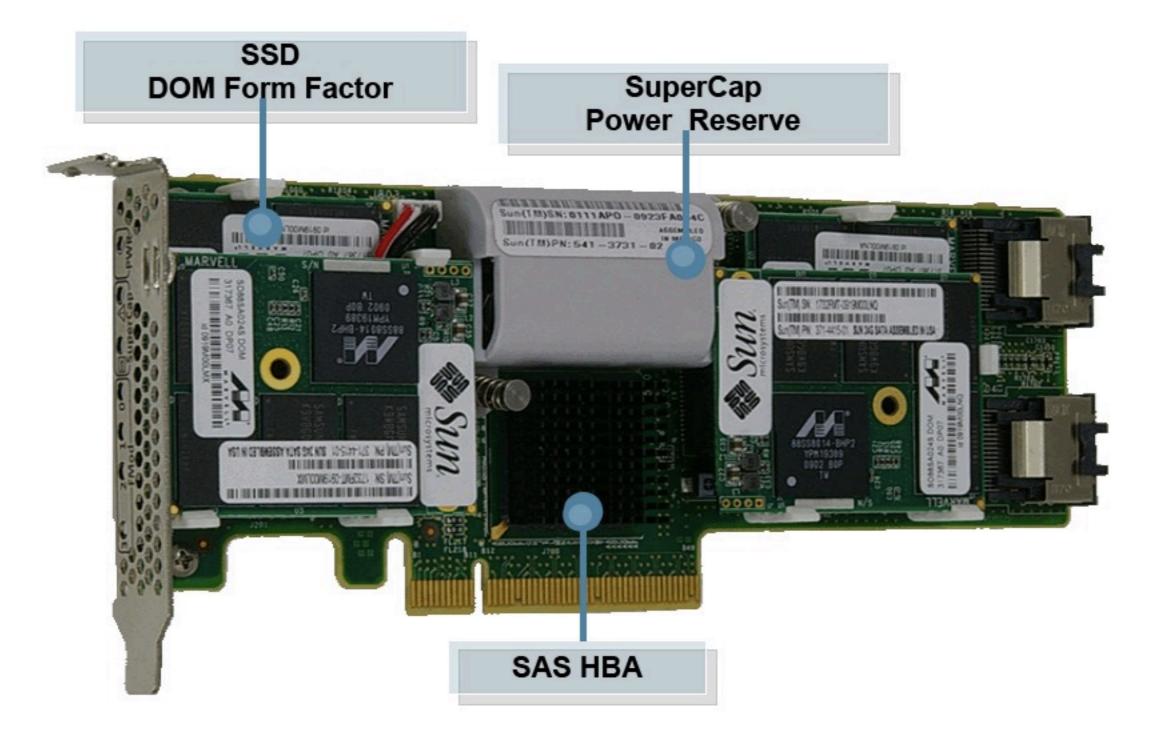


	Yahoo* 40*1U	Blade 6000 with T2 blade	T5240	T5440+J4400
Size	40*1U	4*10U	20*2U	5+5x4U
Thread/Node	8	64	128	256
Disks/Node	4	4	16	24
Memory/Node	8-16 GB	128 GB	256 GB	512 GB
Nodes/Rack	40	40	20*2U	5
Threads/Rack	320	2560	2560	1280
Memory/Rack	320-640 GB	5120 GB	5120 GB	2560 GB
Disks/Rack	160	160	2320	120



### More density? More performance?





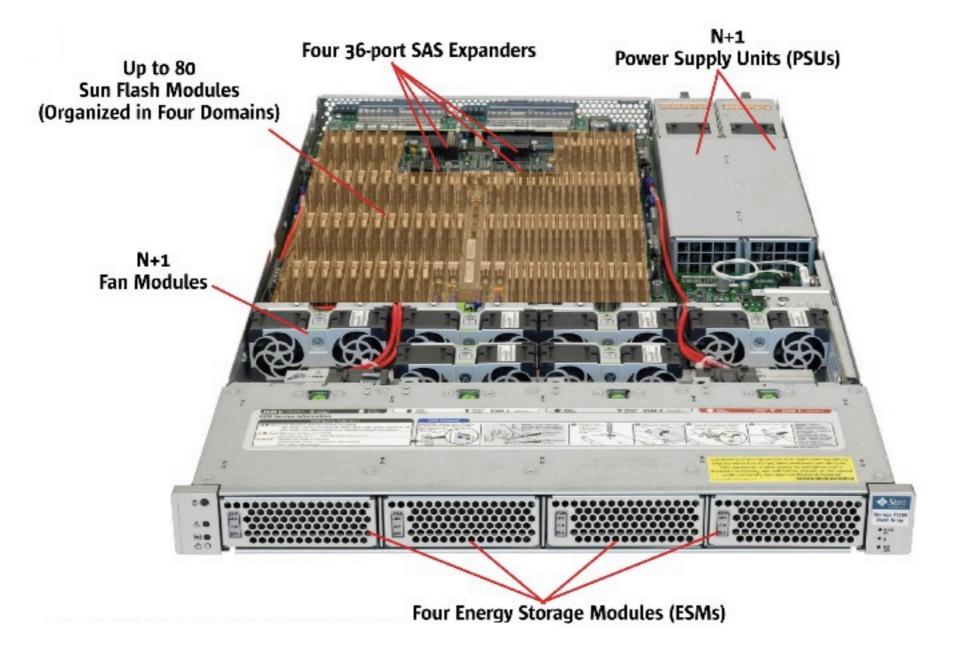


Capacity		
	48 GB	96 GB
Domains	2	4
Flash Type (NAND)	Single Level Cell (SLC)	Single Level Cell (SLC)
Performance[1]		
	48 GB	96 GB
Random Read IOPS (4 K)	53 K IOPS	100 K IOPS
Random Write IOPS (4 K)	42 K IOPS	87 K IOPS
Sequential Read Rate (1024 K transfer size)	546 MB/s	1092 MB/s
Sequential Write Rate (1024 K transfer size)	250 MB/s	494 MB/s



### When you want to go really extreme ... Sun Storage Flash Array F5100







# rack unit million IOPS random write million IOPS random read GByte/s sequential read GByte/s sequential write TB capacity



	Yahoo* 40*1U	Blade 6000	T5240	T5440+J4 400	T5440+F5100	T5120+F5100
Size	40*1U	4*10U	20*2U	5+5x4U	8*(1U + 4U))	20*(1U+1U)
Thread/ Node	8	64	128	256	256	128
Disks/ Node	4	4	16	24	80	80
Memory/ Node	8-16 GB	128 GB	256 GB	512 GB	512	256
Nodes/ Rack	40	40	20	5	8	20
Threads/ Rack	320	2560	2560	1280	2.048	2560
Memory/ Rack	320-64 0	5120 GB	5120 GB	2560 GB	4.096	5120
Disks/ Rack	160	160	320	120	640	1600

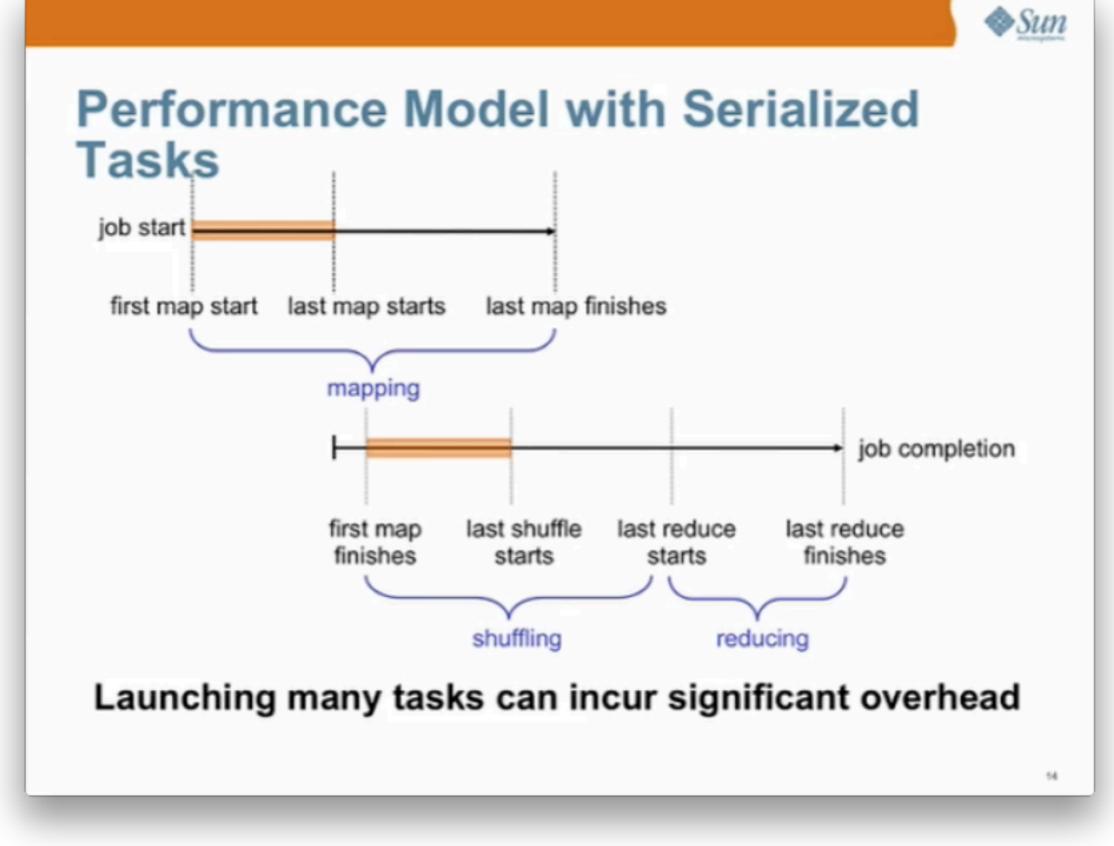


### But colleagues found a problem with such large cluster

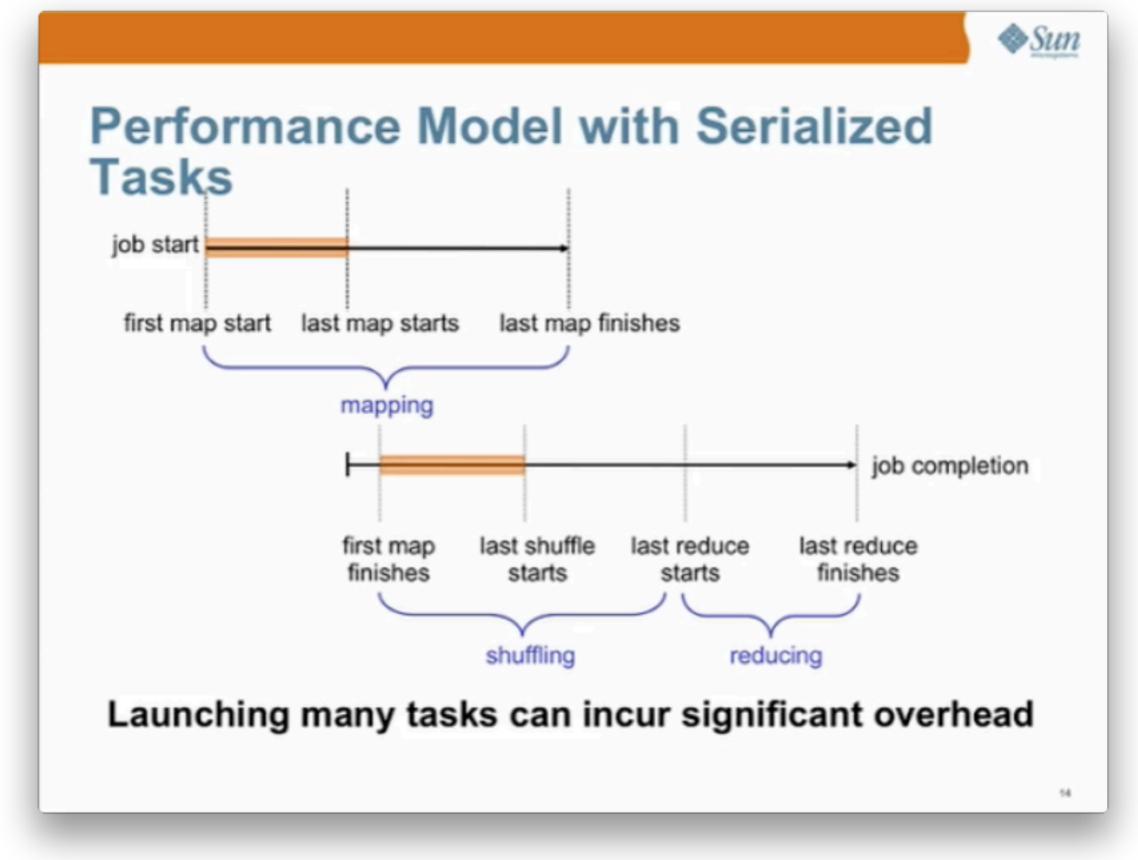


### I will just use their slides now ...

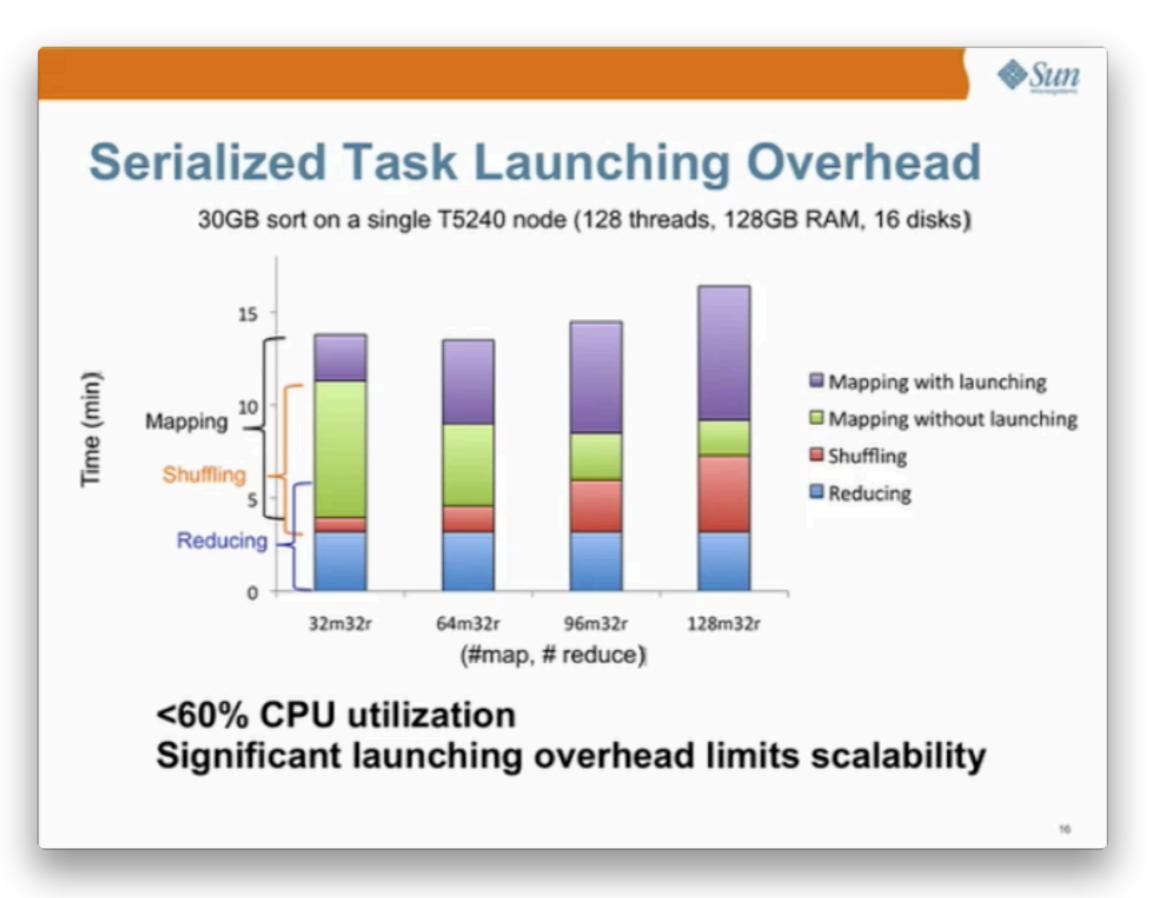




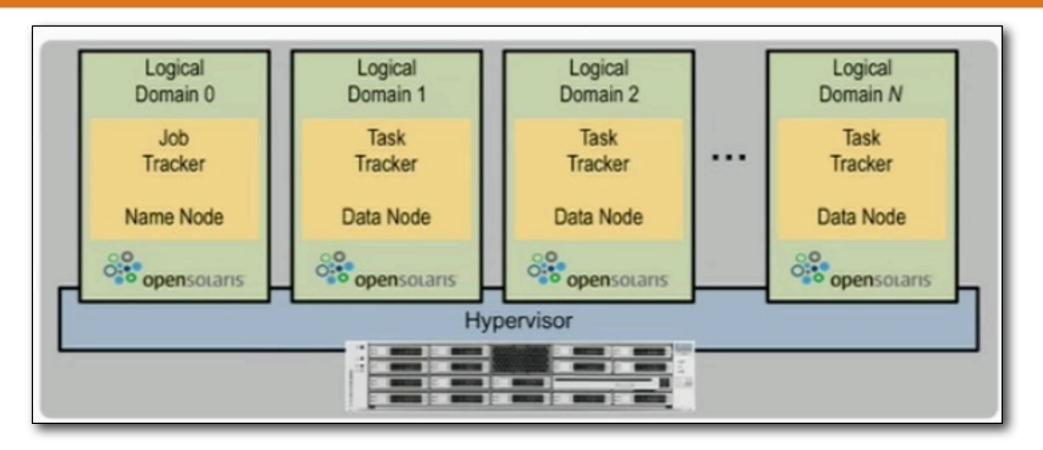


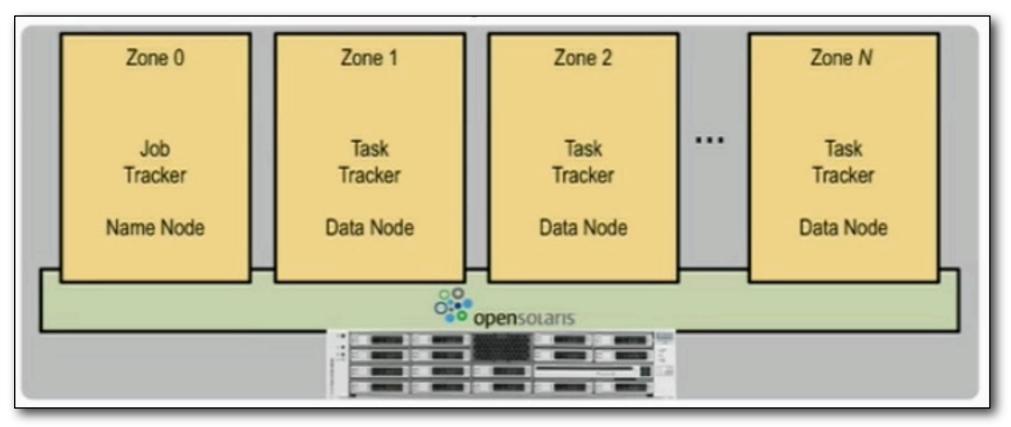








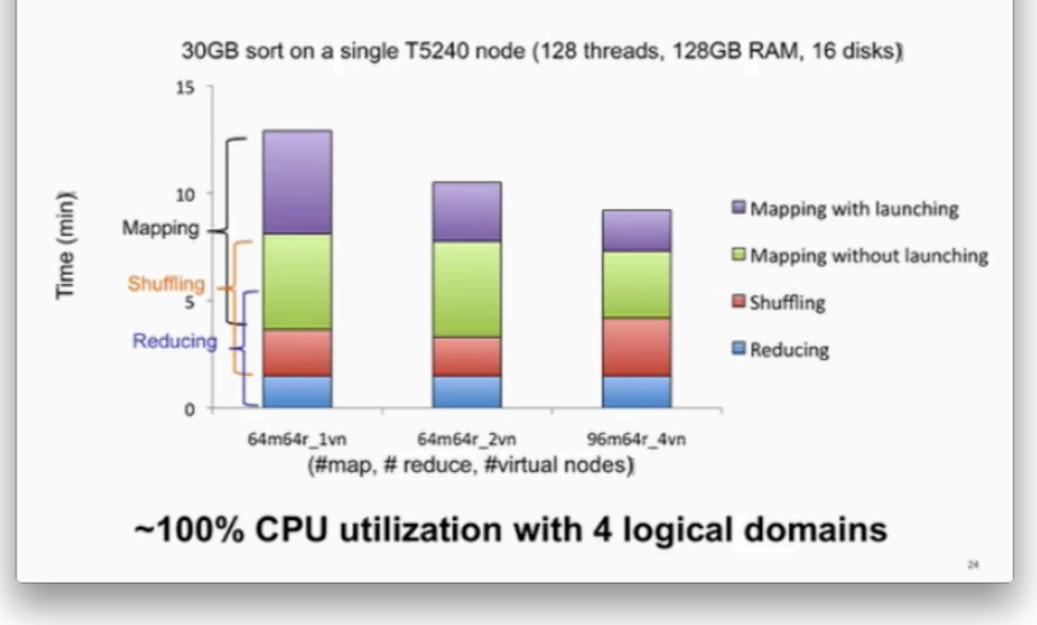




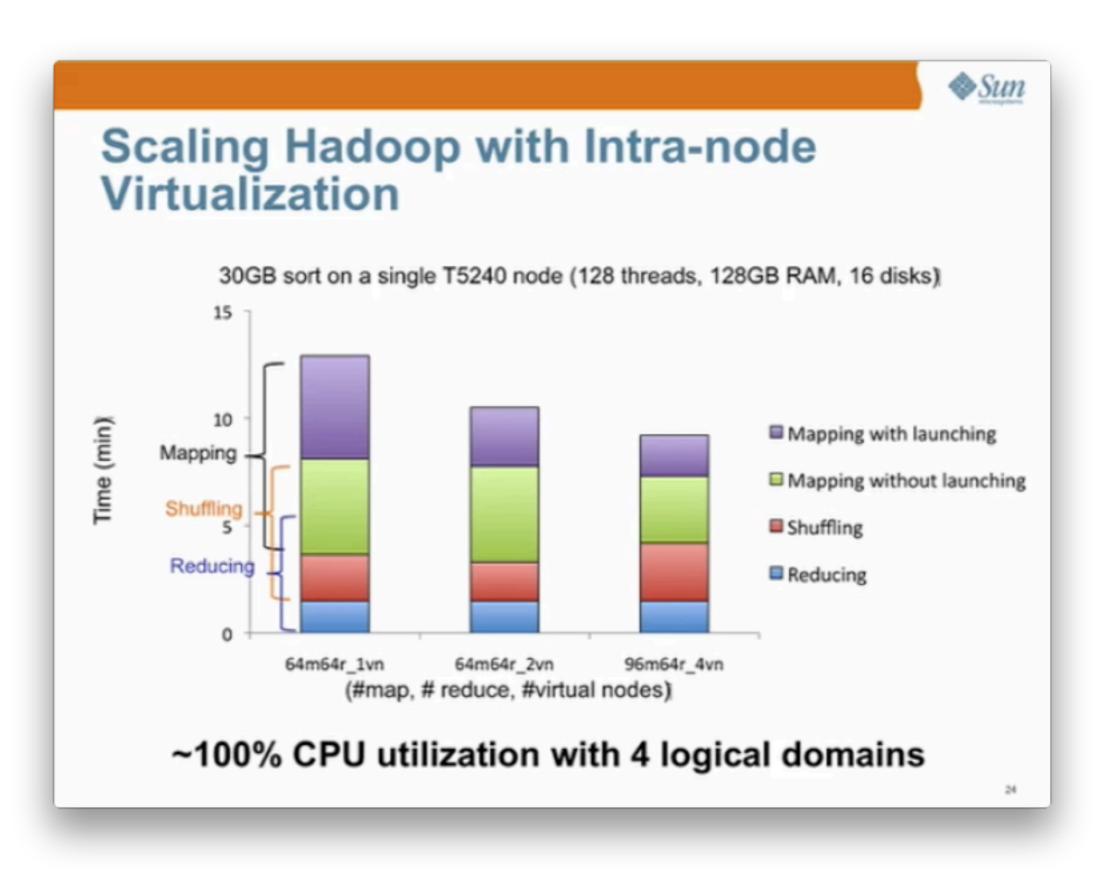


🔷 Sun











#### *♦<u>Sun</u>*

### **E-mail Discovery Overview**

- Preparing data for searching over large email corpus
- Five phases with different MapReduce profiles
  - PipelineMapReduce Reads and parses 27GB of raw emails
  - DocumentSeqFileToMapFile Prepares MapFile to retrieve data
  - PersonNormalization Groups data into unique entities
  - Consumer Creates indices
  - ThreadDetection Conversation threads detected
- Output is a set of shards used in an E-mail discovery search application



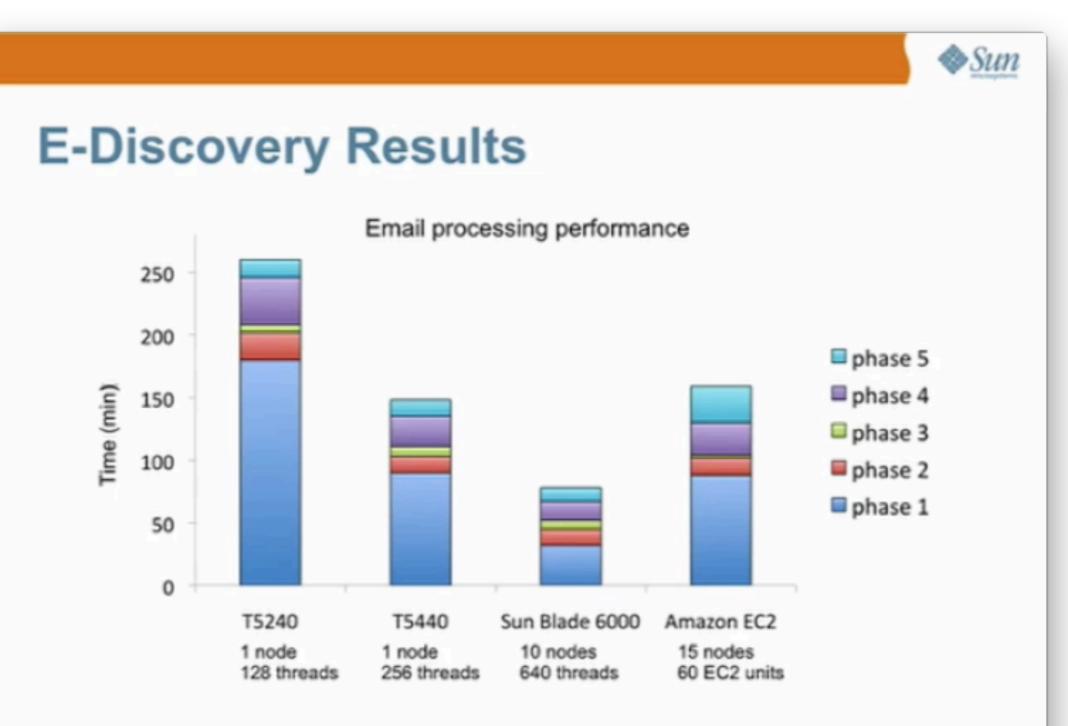
### E-mail Discovery (http://www.it-discovery.com/)

a aga	8, 2008 Through War-28, 2001 1	unana Chan.		dentry (# MAC)	an Bathan ()
80×15,10		Lunial	uluka Pi		
_	eesh. Contension, Recording	Annth   Brown	Dise: beleatings	-	Consectoria
1 Mar	etulita Lastica, Drancationa bauer fitas		atur factor (the	0-1	attes abertera diente an
il be	sa Non adeartos termoritator nexe Develumente Management, Org mis Legal, Financial, Business Challers		2 B af associations and the 2 area as feedback and the manufeedback and the manufeedback and the second second		S estumentarian S estra charactioner con
This Po	age. All. Name ( Non-Housever, ) ( Non-		Actions 1	i)	1 - 12 of SM documents
10	an Al Note Technologie And	3	Super		1 - 10 of 504 documents Data V Salawice
a Dab	Tops singe barrents genor con	Te argan phrastigance con	Subati In Follow Up on Houston Ope		1 - 12 of 594 documents
This Po	an Al Note Technologie And	36. argan ahriaci@arcon.com a tals to Snent Talsal astoh har in the sara apedeca@arron.com	Subset re: Follow Up on Houseon Opp reaming to review w. re: Research and Developmen	aranta	1 - 10 of 504 documents Data V Salawice
2 Dia 3	Anna Ali, Noral Interferences (Inter Engin since karrents@enter.com Argan, fisce another meeting, Press since karringk@enter.com	To prove photogeneous con a telle to Grant I selad sector han in the serie apodecid@amore.com receptations: The connection: Viriae 2003 plant@mather.org.edu.as	Subset rs. Follow (y) on Howeton Opp reprints to involve is. re. Research and Developmen 0000-00-25 PM V. re. my final draft	aranta	1 - 12 of SM documents Data V Bellington 18-19-2012
	All Note Technology Report Page Virtue Aeronal Genon cont Argon, Trace profile meeting, Press Vires Vis etcal gal to the eccentring p virtue karrinal Genon cont	To prove other sectors and a test to Sever. I value and the inter- ways appropriate approximation the ways approximation. Viron 2003 plant@mathies up advues plant@mathies up advues participations with a request to invite participation of the sectors invite participation of the sectors.	Subset rs Foldow (yp an Houston Opp naming to moles a. rs Research and Developme 2000 00:25 Me V. rs my finit draft ryte for an Hotebuch. rs MSOF Speaker Series.	aranta	1 - 10 of 10H decuments Data V Betweener 14-10-2000 08-10-2000
• Dia.N	All Some Terriferences Reserved Free strop karrents@erren.com strop karrents@erren.com strop karrinsk@erren.com Strop karrinsk@erren.com Suerdn, 1 forwerded your resume is o strop karrinsk@erren.com	Te argan atmac@ansor.com a tati to Grant. I atab optimizer ten vers apotecid@ansor.tom te gen@methe up.etu.as or fighting effore with a request to indo permana@outeetu.com in Bakey to aak ter to move her visit to verge kansinasi@prote.com	Subset rs Foldow (yp an Houston Opp naming to moles a. rs Research and Developme 2000 00:25 Me V. rs my finit draft ryte for an Hotebuch. rs MSOF Speaker Series.	aranta	1-10 of 1914 decuments     2015 ▼ Bellowcar     04~10/2000     04~10/2000     06~11/2000
	All, Nore Technology Reserves      Enge      Insex kernels(genon-con      Arpen, Trace another meeting. Please      alnox kernels(genon-con      texe tex and tex te eccounting.g      ence kernels(genon-con      final text), Texe constant to o      inso kernels(genon-con      final text), Texe constant to o      inso kernels(genon-con      final text), Texe constant Allow      resh heading. Texe constant Allow      resh heading. Texe constant Allow      resh heading.texer con      final text, tex specialized for      inso kernels(genon-con      final text), text tex specialized for      inso kernels(genon-con      final text)	To argan atmact@arror con a tab to Grant I shall calcoline in the west apolitica@arror con resp atout the connection. Vives 2009 plant@maths up, etc. as or Sphery office with a request to inde gonwan@oubject.com in Balley to ask for to nove for visit to unce services@oubject.com by plaup. Mark. colder@roce.etc.	Subset In: Follow Up on However Operating to review w. In: Research and Developmen cost co.or of your Priv. In: my first draft ryok for an instance. In: MSOF Speaker Series. Re cartopus to co. In: An interesting resume In: Serienz Series Mag	aranta	1-12 of 1914 decuments     2010 ▼ Researce     06110/2000     06111/2000     06111/2000     20111/2000
	All, Nore Mechanism New Property of the Annual System Control Contr	No priper phranciperson con a table to Grant I tablet calcoline in the sens aproduce@priver com mage about the connection. Vives (2009 diver@mather.org.edu.ex or Spoling office with a request to invite permanage/oper com in balance to move her visit to unce searching@price com my price. Mark. calcolor@price ado eleage. Virise Barkers Deates read primerarean sinstranea@price	Subset rs: Follow Vp on However Oper rearrang to review w re: Research and Developmen 0000-00-pic PMr V re: my finit draft synx for an instatut re: MSOF Speaker Series the campus to tol re: An interesting resume rs: Benning Terms Mug et@nice.abcr rs: V50-face.	aranta	1-12 of 1914 decuments     06 110 2000     06 110 2000     06 110 2000     06 110 2000     06 110 2000     06 110 2000     06 110 2000     06 110 2000
0 0 0 0 0	All, Nore Mechanism Responses     Train     Insue     Insue	No priper phranciperson con a table to Grant I tablet calcoline in the sens aproduce@priver com mage about the connection. Vives (2009 diver@mather.org.edu.ex or Spoling office with a request to invite permanage/oper com in balance to move her visit to unce searching@price com my price. Mark. calcolor@price ado eleage. Virise Barkers Deates read primerarean sinstranea@price	Subset rs: Follow Vp on However Oper rearrang to review w re: Research and Developmen 0000-00-pic PMr V re: my finit draft synx for an instatut re: MSOF Speaker Series the campus to tol re: An interesting resume rs: Benning Terms Mug et@nice.abcr rs: V50-face.	aranta	1-10 df 1914 decuments     2010 ▼ Retractat     08-10-2000     08-10-2000     08-11-2000     08-11-2000     08-14-2000     08-15-2000
	All, Nore Technology (     All Store     Free      Strok kerneski@erren.com     Argen, 11-ass andher neeting. Please     sings kerneski@erren.com     Sing karninski@erren.com     Suerie, 11 forwerteld your resume to o     vings kerneski@erren.com     Perre Philips, 1 have contacted Aller     rungs kerneski@erren.com     Freite Heure, bit too specialized for i     vings kerneski@erren.com     Senters, 11 tools great.the server its     Senters, 11 tools great.the server its     finature, bit too specialized for i     vings kerneski@erren.com     Senters, 11 tools great.the server its     finature, bit too specialized for i     vings kerneski@erren.com     Senters, 11 tools great.the server its     finature, bit too specialized for i     vings kerneski@erren.com	No argan atmaci@arror con a sati to Grant i shall open house the vers aprotoco@arror con may about ho connection. Vives 2003 given@mithe up obusis in fighting office with a request to indo commit@indo.con in Balay to ask for to nove her visit to uncut contracting order con in googe Mack colone@rise ado original. Vives Barbara Deater read prevenances in the Research Group. If gives investing women cont is.Res_contracting women cont is.Res_contracting women cont is.Res_contracting women cont	Subset In: Following on Houseon Operating to molecule and reprinting to molecule ac. In: Research and Development total applications In: Research and Development reprint dust reprint dust reprint dust In: Million State In: Semanar Series Mug stightion adurt In: Research Researce In: Followers	aranta	1-12 df 1914 decuments      Dats ▼     Datsucat      De1=02000      De1=020

27

**♦**<u>Sun</u>

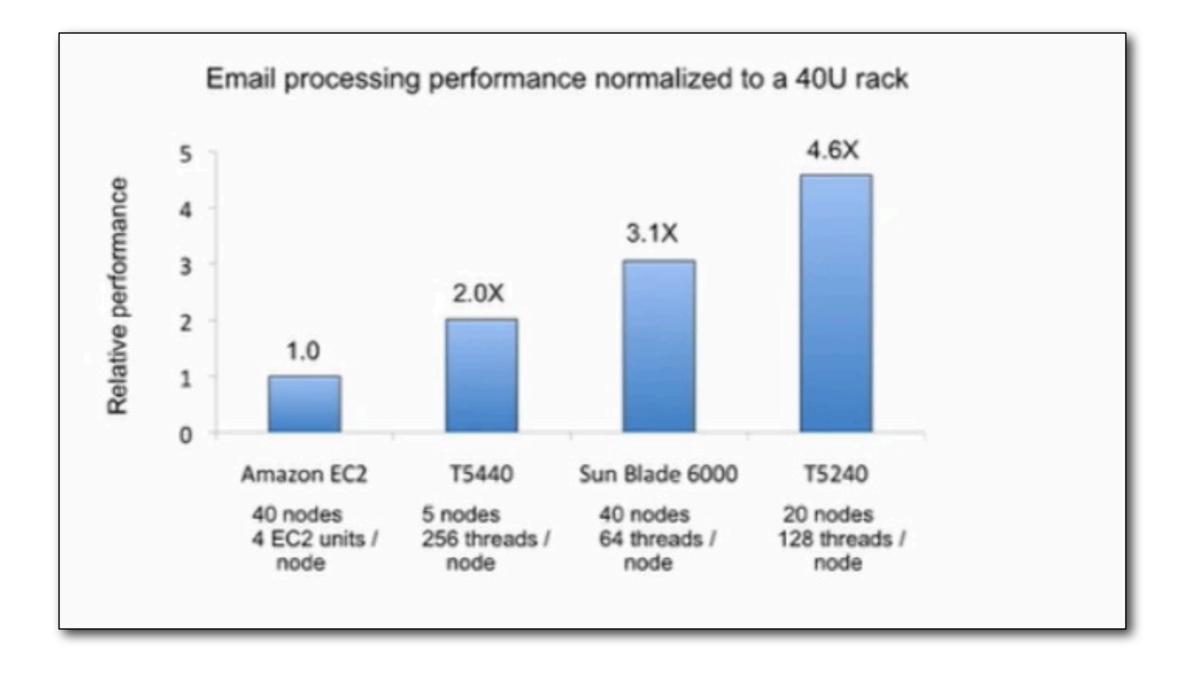




CMT Hadoop systems scale for throughput applications

28







### Solaris+Hadoop



### I've already talked about Logical Domains and Zones



### There is a build-in virtualization in Solaris It's called Zones.



## It's an low/no-overhead virtualization



a single kernel look as several ones.



Thus you have a virtual operating system in your OS.



Up to 8191.



## ... you will have no memory before reaching this number.



#### A Solaris Zone

... can't access the hardware directly ... has it's own root ... can't see the contents of other zones ... is a resource management entity



### So you could use your normal server systems.

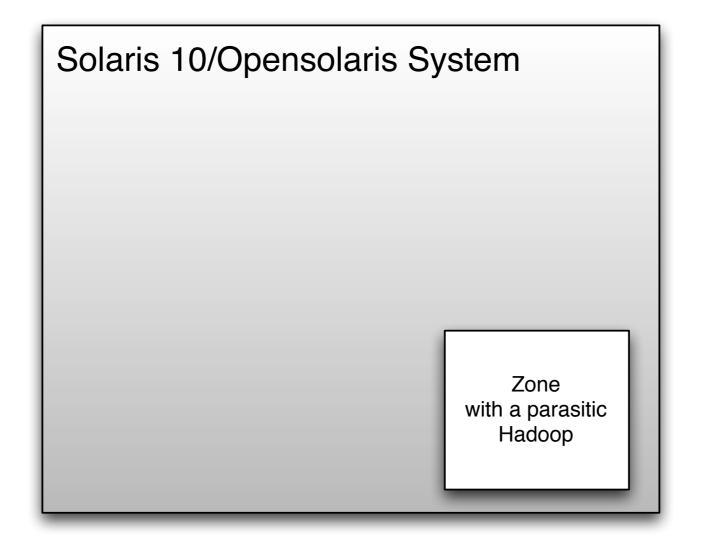


#### Parasitic Hadoop



#### It lives from the idle cycles on your systems.







#### A parasite has to ensure that it doesn't kill the host, as it would kill the parasite as well.



## Solaris has a functionality called Solaris Resource Management



You can limit the consumption: ... of CPU cycles ... of memory consumption ... of swap space ... of network bandwith



## #! /usr/bin/perl while (1) { my \$res = ( 3.3333 / 3.14 ) }



#### # su einstein

#### Password:

- \$ /opt/bombs/cpuhog.pl &
- \$ /opt/bombs/cpuhog.pl &



#### bash -3.2\$ ps-opcpu ,project ,args %CPU PROJECT COMMAND 0.0 user.einstein -sh

- 0.3 user.einstein bash
- 47.3 user.einstein /usr/bin/perl /opt/bombs/cpuhog.pl
- 48.0 user.einstein /usr/bin/perl /opt/bombs/cpuhog.pl
- 0.2 user.einstein ps -o pcpu, project, args



#### # dispadmin -d FSS



## # projadd shcproject # projmod -U einstein shcproject

# projmod -K "project.cpu-shares=(privileged ,150,none)" lhcproject
# projmod -K "project.cpu-shares=(privileged ,50,none)" shcproject



\$ newtask -p shcproject /opt/bombs/cpuhog.pl &

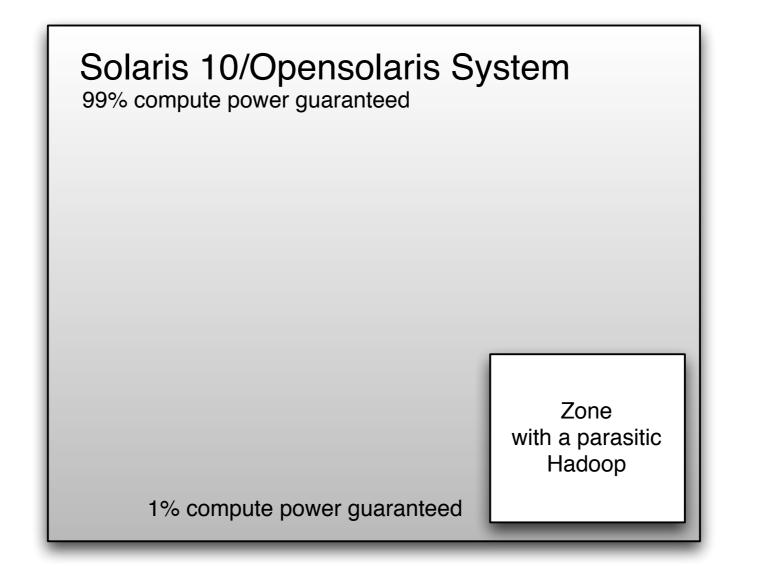
```
$ ps -o pcpu ,project ,args
%CPU PROJECT COMMAND
0.0 user.einstein -sh
0.3 user.einstein bash
0.2 user.einstein ps -o pcpu,project,args
95.9 shcproject /usr/bin/perl /opt/bombs/cpuhog.pl
```



```
$ newtask -p lhcproject /opt/bombs/cpuhog.pl &
[2] 784
```

```
$ ps -o pcpu ,project ,args
%CPU PROJECT COMMAND
0.0 user.einstein -sh
0.1 user.einstein bash
72.5 lhcproject /usr/bin/perl /opt/bombs/cpuhog.pl
25.6 shcproject /usr/bin/perl /opt/bombs/cpuhog.pl
0.2 user.einstein ps -o pcpu,project,args
```







# Icing on the cake ZFS



#### Forget everything you know about filesystems: ZFS isn't really a filesystem ... A POSIX compatible filesystem is just a possible view an emulated block device is another ...



#### No volumes Integrated RAID (RAID done right - RAID5/RAID6/RAID TP without read-amplification and write-hole) Usage-aware selective resilvering Creating filesystem as easy as directories Guaranteed data validity (okay 99,999999999999999999999) Guaranteed consistent on-disk state of the filesystem Integrated compression Integrated Deduplication



#### More important for our "parasitic Hadoop": Quota+Reservations

#### Putting the HDFS in an own filesystem

Reservation: ensuring that a filesystem has a certain minimum of free space that can't be used by other filesystems

Quota:

ensuring that a filesystem can't get bigger than a certain size.



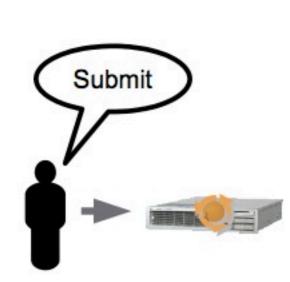
## Sun Grid Engine+Hadoop

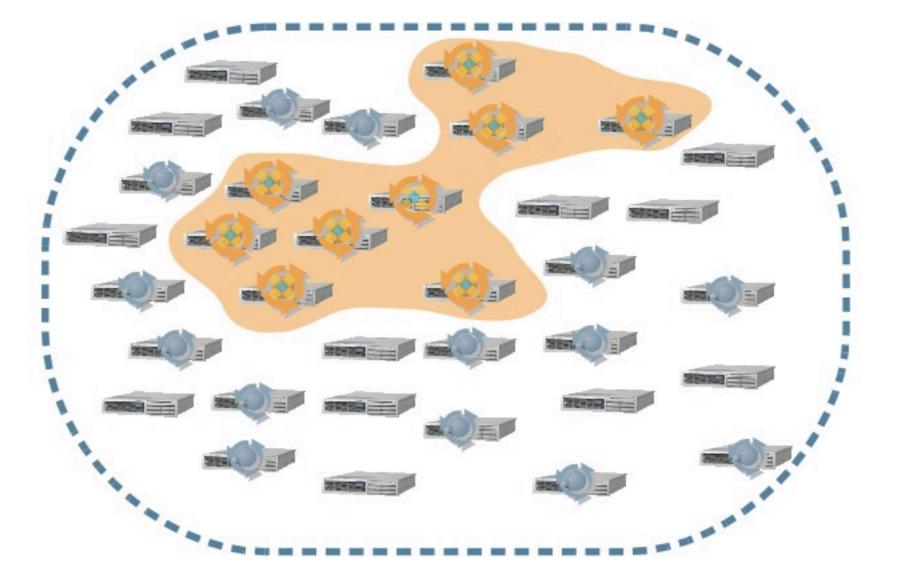


Great by itself on dedicated machines Map/reduce only Unaware of other machine load Schedules only against data No policies No resource differentiation No accounting

All things that DRMs do well

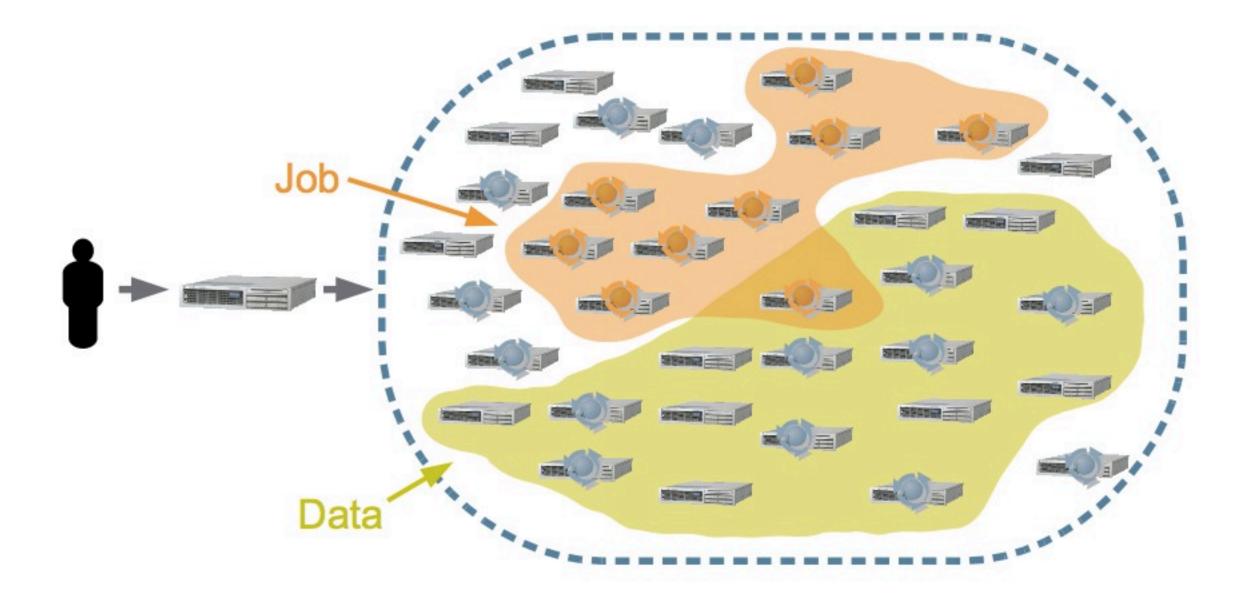








#### The Hadoop-on-Demand works resonably well but has a problem: It doesn't know about the location of the data in the HDFS.





#### **Scheduling Against the Data**

#### Grid Engine resources, aka "complexes"

#### Model aspects of your cluster

Concrete Free memory Software licenses Abstract High priority Exclusive host access Can be fixed, counted, or *measured* 

#### Why not model HDFS data blocks as resources?

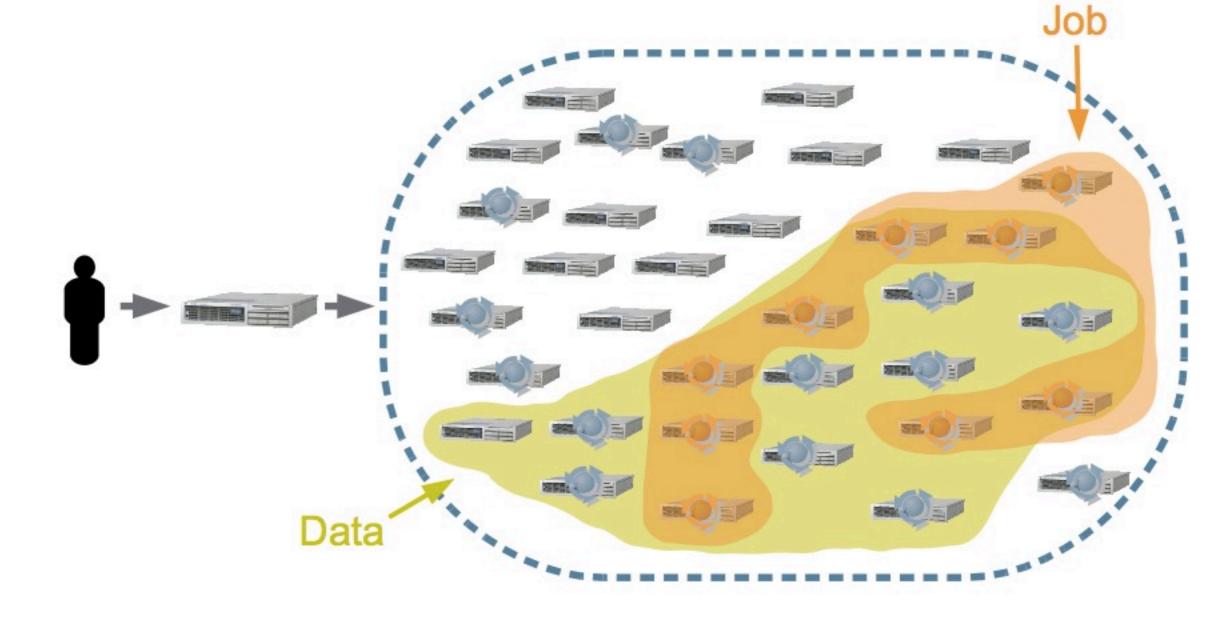


#### **Scheduling Against the Data**

The new integreation "measures" where blocks are ... ... a helper software finds out which blocks you need ... ... and schedules your Hadoop accordingly on this grid nodes.



## The new Sun Grid Engine integration of hadoop is data locality aware





# Vielen Dank für Ihre Aufmerksamkeit!

Jörg Möllenkamp Principal Field Technologist

**Sun Microsystems**