Farms, Fabrics and Clouds

Steve Loughran
Julio Guijarro
HP Laboratories, Bristol, UK

December 2007

steve.loughran@hpl.hp.com
julio.guijarro@hpl.hp.com

Julio Guijarro

Researcher at HP Laboratories

Area of interest: Deployment

In charge of OSS release

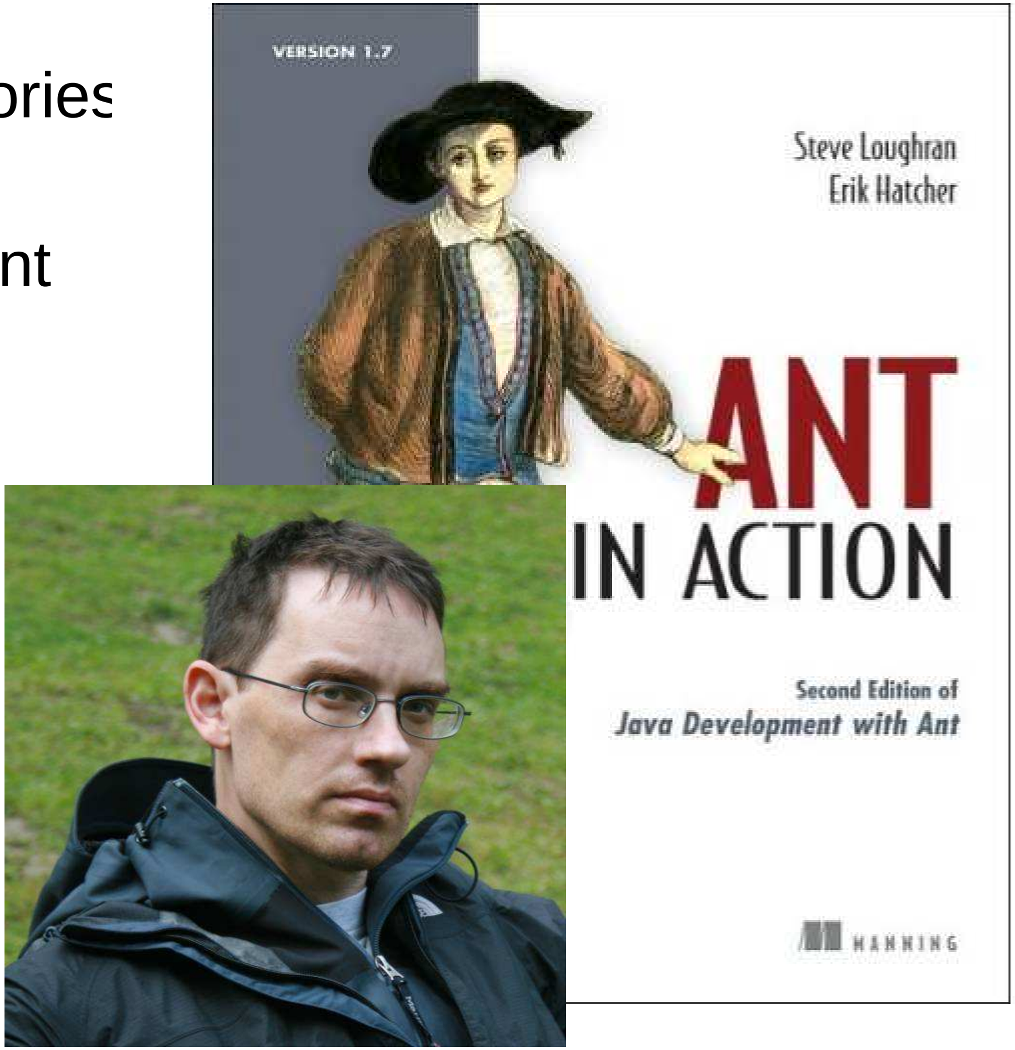
http://smartfrog.org/

Steve Loughran

Researcher at HP Laboratories

Area of interest: Deployment

Author of *Ant in Action*



Our research

- How to host big applications across distributed resources
  - Automatically
  - Repeatably
  - Dynamically
  - Correctly
  - Securely
- How to manage them from installation to removal
- How to make dynamically allocated servers useful

Question

Who had breakfast this morning?

Question

Who harvested wheat or corn,
or killed an animal
for
that breakfast?

Farms provide food.

It is *somebody else's problem*

Question

Who is wearing clothes they wove or knitted themselves?

Provisioning of clothing -fabrics- is outsourced

It is *somebody else's problem*

All new applications are on the Web

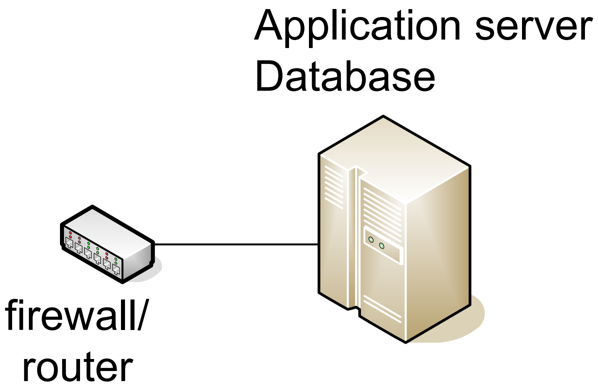
- Web Browser, AJAX clients
- Richer: Flash, XUL, Silverlight
- "… as a Service "

⇒Lots of code running in the server

⇒Data mining/analysis problems

⇒Unpredictable demand

Old world installation: single server



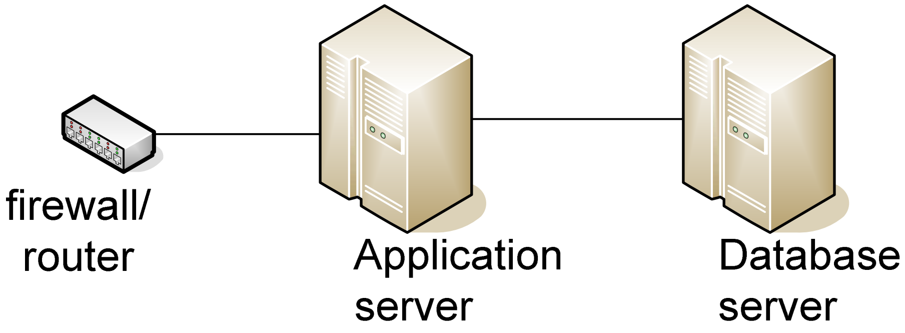
Application server
Database

firewall/
router

Single web server,

Single DB

RAID filestore



firewall/
router

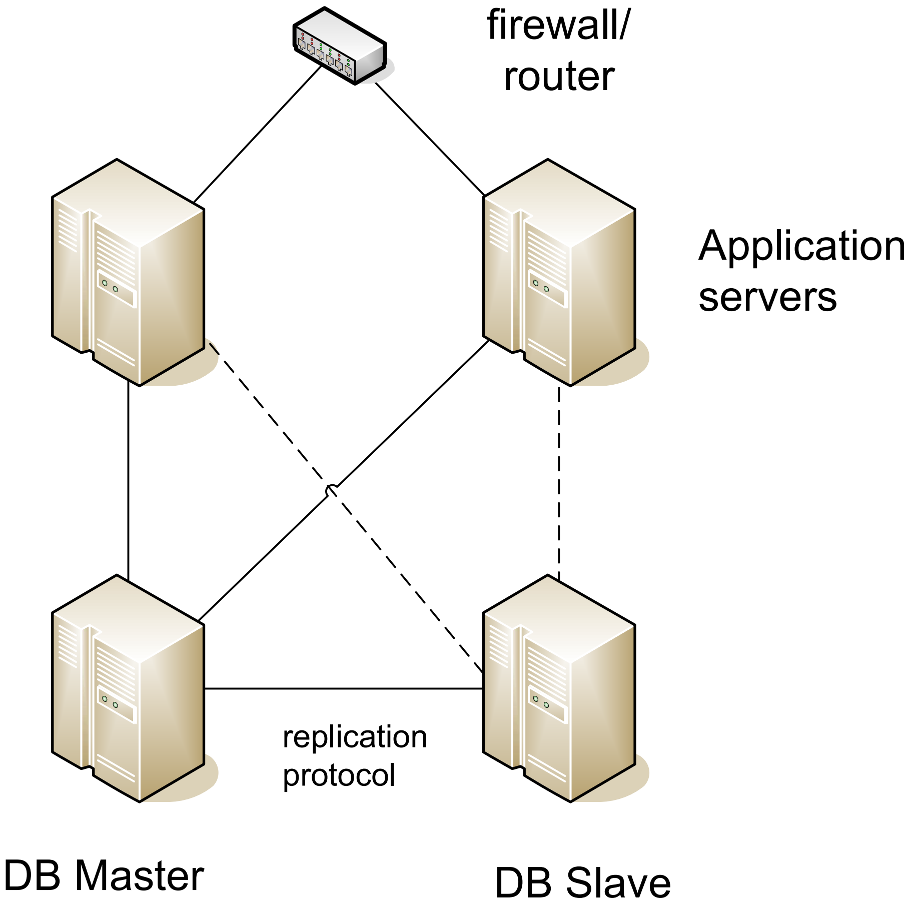
Application
server

Database
server

-SPOF

-limitations of scale

yesterday: clustering



Multiple web servers,

Replicated DB

RAID Network filestore

Load-balancing router

-Cost

-Complexity

-Limitations of scale

Maintains the illusion of a single server

Now: server farms

500 web servers,

Distributed filestore

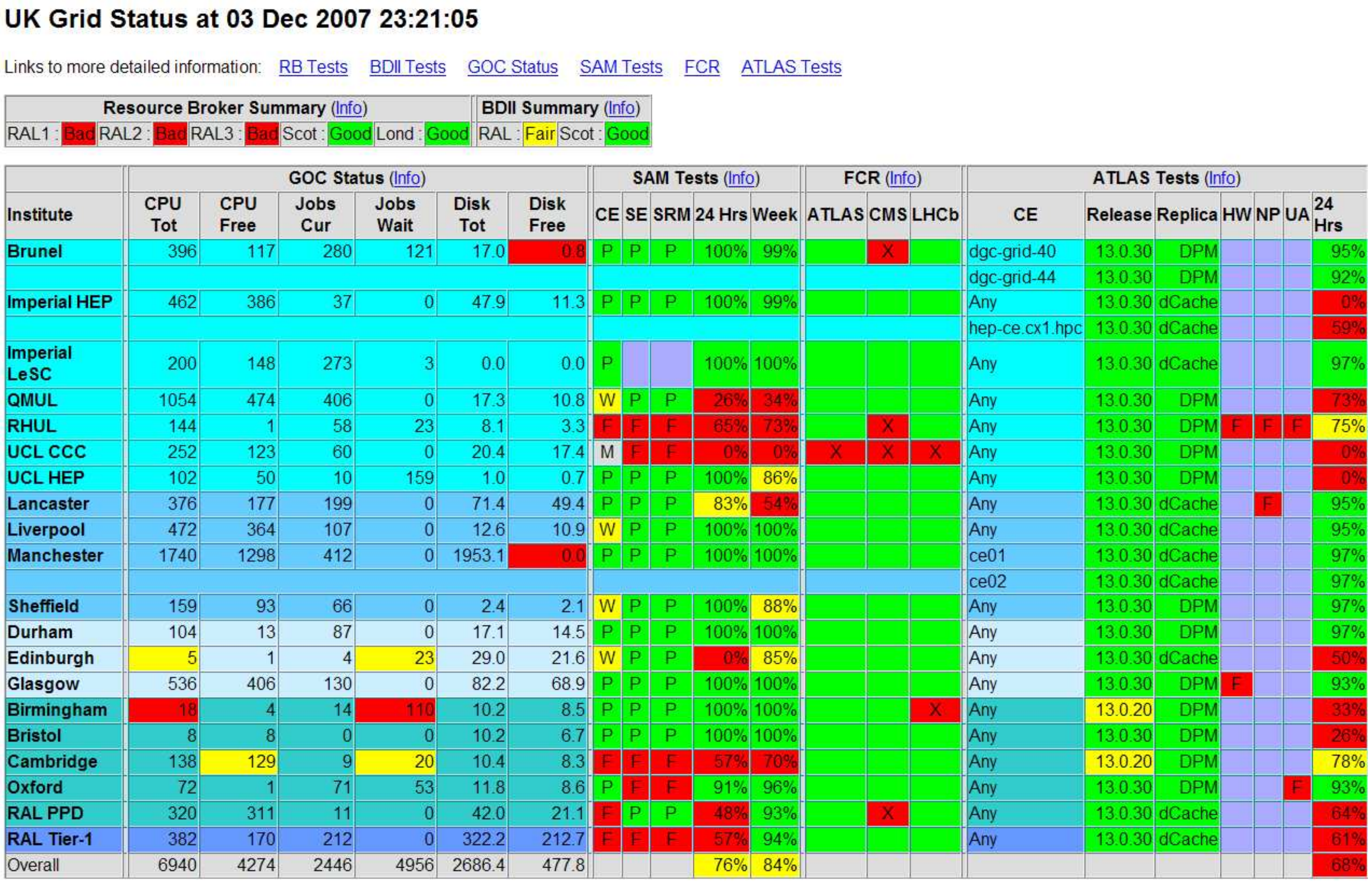
Rented storage & CPU

Scales up

No capital outlay

Agile infrastructure

tomorrow? grid fabric. 50000 servers



UK Grid Status at 03 Dec 2007 23:21:05

Links to more detailed information: RB Tests BDII Tests GOC Status SAM Tests FCR ATLAS Tests

Resource Broker Summary (Info)					BDII Summary (Info)	
RAL1 : Bad	RAL2 : Bad	RAL3 : Bad	Scot : Good	Lond : Good	RAL : Fair	Scot : Good

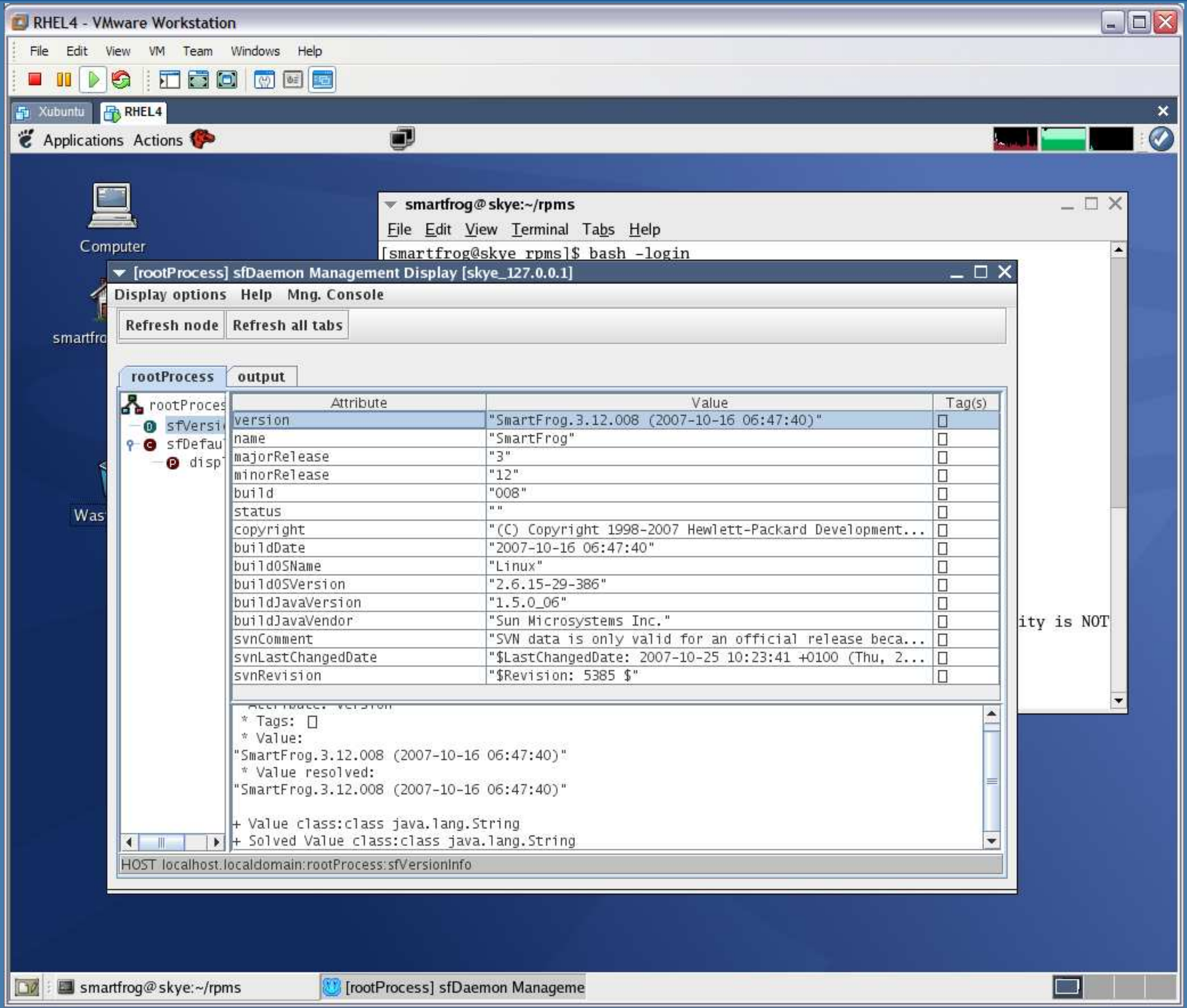
	GOC Status (Info)						SAM Tests (Info)					FCR (Info)			ATLAS Tests (Info)						
Institute	CPU Tot	CPU Free	Jobs Cur	Jobs Wait	Disk Tot	Disk Free	CE	SE	SRM	24 Hrs	Week	ATLAS	CMS	LHCb	CE	Release	Replica	HW	NP	UA	24 Hrs
Brunel	396	117	280	121	17.0	0.8	P	P	P	100%	99%		X		dgc-grid-40	13.0.30	DPM				95%
															dgc-grid-44	13.0.30	DPM				92%
Imperial HEP	462	386	37	0	47.9	11.3	P	P	P	100%	99%				Any	13.0.30	dCache				0%
															hep-ce.cx1.hpc	13.0.30	dCache				59%
Imperial LeSC	200	148	273	3	0.0	0.0	P			100%	100%				Any	13.0.30	dCache				97%
QMUL	1054	474	406	0	17.3	10.8	W	P	P	26%	34%				Any	13.0.30	DPM				73%
RHUL	144	1	58	23	8.1	3.3	F	F	F	65%	73%		X		Any	13.0.30	DPM	F	F	F	75%
UCL CCC	252	123	60	0	20.4	17.4	M	F	F	0%	0%	X	X	X	Any	13.0.30	DPM				0%
UCL HEP	102	50	10	159	1.0	0.7	P	P	P	100%	86%				Any	13.0.30	DPM				0%
Lancaster	376	177	199	0	71.4	49.4	P	P	P	83%	54%				Any	13.0.30	dCache		F		95%
Liverpool	472	364	107	0	12.6	10.9	W	P	P	100%	100%				Any	13.0.30	dCache				95%
Manchester	1740	1298	412	0	1953.1	0.0	P	P	P	100%	100%				ce01	13.0.30	dCache				97%
															ce02	13.0.30	dCache				97%
Sheffield	159	93	66	0	2.4	2.1	W	P	P	100%	88%				Any	13.0.30	DPM				97%
Durham	104	13	87	0	17.1	14.5	P	P	P	100%	100%				Any	13.0.30	DPM				97%
Edinburgh	5	1	4	23	29.0	21.6	W	P	P	0%	85%				Any	13.0.30	dCache				50%
Glasgow	536	406	130	0	82.2	68.9	P	P	P	100%	100%				Any	13.0.30	DPM	F			93%
Birmingham	18	4	14	110	10.2	8.5	P	P	P	100%	100%			X	Any	13.0.20	DPM				33%
Bristol	8	8	0	0	10.2	6.7	P	P	P	100%	100%				Any	13.0.30	DPM				26%
Cambridge	138	129	9	20	10.4	8.3	F	F	F	57%	70%				Any	13.0.20	DPM				78%
Oxford	72	1	71	53	11.8	8.6	P	F	F	91%	96%				Any	13.0.30	DPM			F	93%
RAL PPD	320	311	11	0	42.0	21.1	F	P	P	48%	93%		X		Any	13.0.30	dCache				64%
RAL Tier-1	382	170	212	0	322.2	212.7	F	F	F	57%	94%				Any	13.0.30	dCache				61%
Overall	6940	4274	2446	4956	2686.4	477.8				76%	84%										68%

Application architectures and deployment problems change radically in this world

Application architectures

- ROA/REST
- Virtualized
- Map/Reduce
- Shards
- Tuple-spaces
- Grid

Virtualization



RHEL4 - VMware Workstation
File Edit View VM Team Windows Help
Xubuntu
RHEL4
Applications Actions
Computer
smartfro
Was
smartfrog@skye:~/rpms
File Edit View Terminal Tabs Help
[rootProcess] sfDaemon Management Display [skye_127.0.0.1]
Display options Help Mng. Console
Refresh node
Refresh all tabs
rootProcess
output
rootProces
sfVersi
sfDefau
disp
smartfrog@skye:~/rpms
[rootProcess] sfDaemon Manageme

```
[smartfrog@skye rpms]$ bash -login
ity is NOT
```

Attribute	Value	Tag(s)
version	"SmartFrog.3.12.008 (2007-10-16 06:47:40)"	□
name	"SmartFrog"	□
majorRelease	"3"	□
minorRelease	"12"	□
build	"008"	□
status	""	□
copyright	"(C) Copyright 1998-2007 Hewlett-Packard Development...	□
buildDate	"2007-10-16 06:47:40"	□
buildOSName	"Linux"	□
buildOSVersion	"2.6.15-29-386"	□
buildJavaVersion	"1.5.0_06"	□
buildJavaVendor	"Sun Microsystems Inc."	□
svnComment	"SVN data is only valid for an official release beca...	□
svnLastChangedDate	"$LastChangedDate: 2007-10-25 10:23:41 +0100 (Thu, 2...	□
svnRevision	"$Revision: 5385 $"	□

```
 * Tags: □
 * Value:
"SmartFrog.3.12.008 (2007-10-16 06:47:40)"
 * Value resolved:
"SmartFrog.3.12.008 (2007-10-16 06:47:40)"

+ Value class:class java.lang.String
+ Solved Value class:class java.lang.String
```

HOST localhost.localdomain:rootProcess:sfVersionInfo

Why?

- Save on hardware (and power, space)
- Dynamically move running servers
- Demand creation of new images
- Testing complex system configurations
- Redistributing entire machine image
- 'virtual appliance'

Assumptions that are now invalid

- Systems have a long lifespan
- It is slow/expensive to create a new system
- It is expensive to duplicate one
- Systems can/should be managed by hand
- Clocks proceed at the same rate
- Physical RAM doesn’t get swapped out
- Running machines can't be moved/cloned

Server Farms

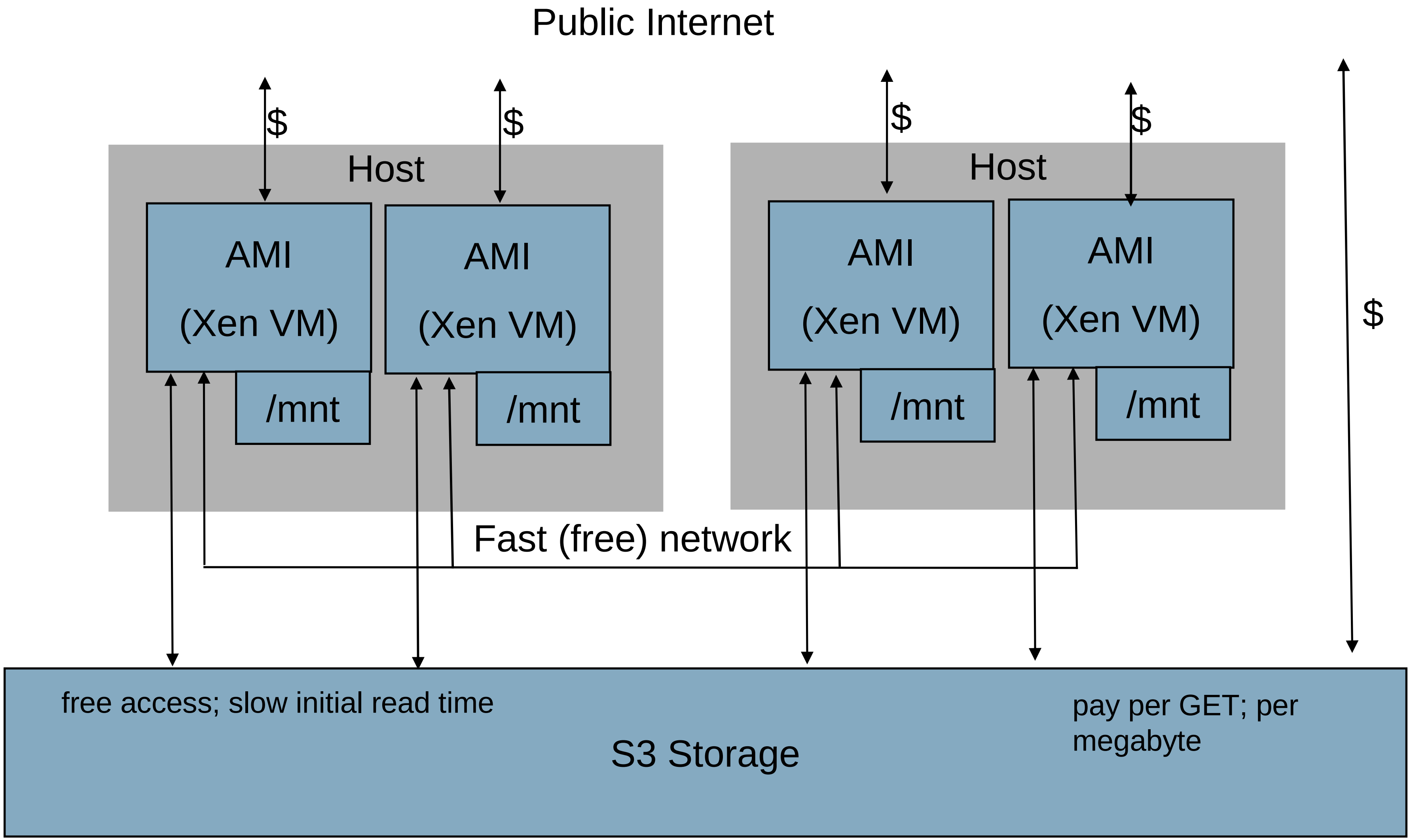
Assumptions that are now invalid

- System failure is an unusual event
- 100% availability can be achieved
- Data is always near the server
- You need physical access to the severs
- Databases are the best storage form
- You need millions of $/£/€ to play

Who has the servers?

- Yahoo, Google, MSN, eBay: services
- MMORPG Game Vendors:
  Word of Warcraft, Second Life
- EU Grid: Scientists
- HP, IBM, Sun: rent to companies
  -focus on CPU performance
- Amazon: rent to anyone with an Amazon account
  -focus on startups

Amazon EC2



Public Internet

Host

AMI

(Xen VM)

/mnt

Fast (free) network

free access; slow initial read time

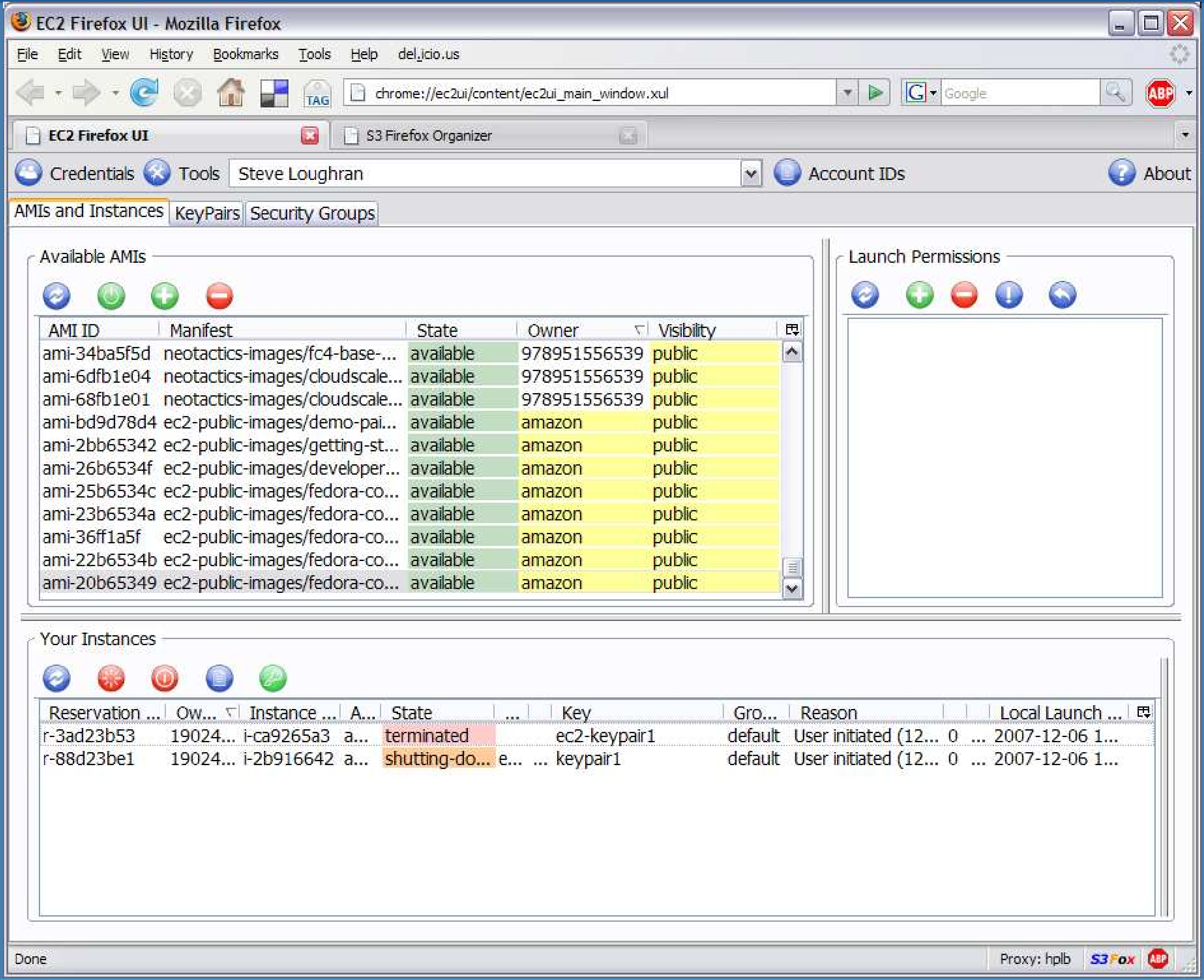
S3 Storage

pay per GET; per megabyte

Amazon EC2

- Pay as you go Virtual Machine Hosting
- No persistent storage other than S3 filestore - uses HTTP GET/PUT/DELETE operations
- $0.10 per CPU/hour
- S3 Storage has own billing
  (by MB & by access -cheaper in bulk)

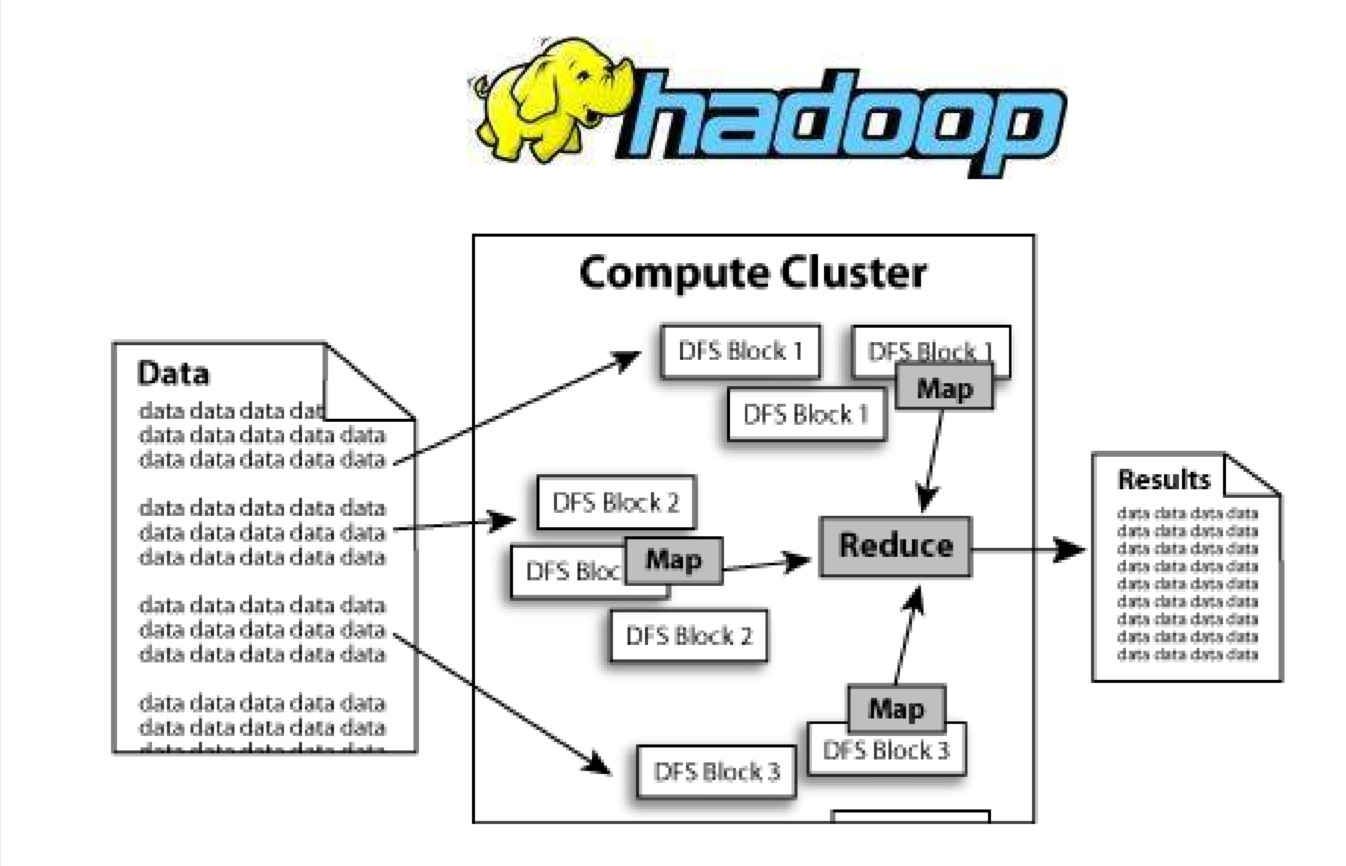
Demo



AMI ID	Manifest	State	Owner	Visibility
ami-34ba5f5d	neotactics-images/fc4-base-...	available	978951556539	public
ami-6dfb1e04	neotactics-images/cloudscale...	available	978951556539	public
ami-68fb1e01	neotactics-images/cloudscale...	available	978951556539	public
ami-bd9d78d4	ec2-public-images/demo-pai...	available	amazon	public
ami-2bb65342	ec2-public-images/getting-st...	available	amazon	public
ami-26b6534f	ec2-public-images/developer...	available	amazon	public
ami-25b6534c	ec2-public-images/fedora-co...	available	amazon	public
ami-23b6534a	ec2-public-images/fedora-co...	available	amazon	public
ami-36ff1a5f	ec2-public-images/fedora-co...	available	amazon	public
ami-22b6534b	ec2-public-images/fedora-co...	available	amazon	public
ami-20b65349	ec2-public-images/fedora-co...	available	amazon	public

Reservation ...	Ow...	Instance ...	A...	State	...		Key	Gro...	Reason			Local Launch ...
r-3ad23b53	19024...	i-ca9265a3	a...	terminated			ec2-keypair1	default	User initiated (12...	0	...	2007-12-06 1...
r-88d23be1	19024...	i-2b916642	a...	shutting-do...	e...	...	keypair1	default	User initiated (12...	0	...	2007-12-06 1...

Map/Reduce



Run code near the data, then merge the results

Assumptions that are now invalid

- Terabyte datasets are hard to work with
- Code runs on a single machine
- Sequential code is better than parallel code
- RAID hardware is the best way to store data
- Databases are better than filesystems

Shards

Assumptions that are now invalid

- A single farm needs to scale to infinity
- You need to provide 100% availability to 100% of users
- You have to roll out simultaneous updates to the application, changes to the DB schema, *globally*

Changes for developers

- Many classic assumptions are invalid

- Design for scale
- Rent servers from the outset
—every developer can have their own set
- Cover your server costs from the outset and you are in the black from day 1

Problems for us farmers

- Power management
- Predictive disk failure management
- Load balancing for availability, power
- Data cache management
- Billing
- Security/Isolation
- How this will change server hardware
- Managing/Configuring Machine Images
- Diagnostics when things go wrong

Topic for discussion

Where is all this heading?