HPC Cloud

Floris Sluiter SARA computing & networking services

Big Grid the dutch e-science grid

About SARA, NCF and BiG Grid

- The foundation for National Compute Facilities is part of NWO, the Dutch Government Organization for Scientific Research
- The **BiG Grid** project is a collaboration between NCF, Nikhef and NBIC, and enables access to grid infrastructures for scientific research in the Netherlands.
- **SARA** is a national High Performance Computing and e-Science Support Center, in Amsterdam and the primary operational partner of BiG Grid











SARA Project involvements



Scientific Infrastructure and support

SARA

High Performance Computing	Huygens, GPU cluster Lisa, Grid. Hadoop, HPC Cloud	
High Resolution Visualization	Tiled Panel Display Remote Visualization	
High Performance Networking	SURFnet 6 AMSix Netherlight	
Mass Storage	2*10 Petabyte Tape archive 4 Petabyte disk storage	
BiG Grid		

the dutch e-science grid

Scientific Computing facilities SARA (Specs)

Huygens National Super

Power6, 3328 cores in 105 nodes 15.25 TB of memory, Infiniband 160 Gbit/s 700 TB of disk space, 60 TFlop/s



LISA National Compute Cluster

Intel, 4480 cores in 512 nodes, 12 TB of memory, Infiniband 20Gbit/s 50Tbyte disk space 20 TFlop/s



Grid Resources

Intel, 3000 Cores in 3000 nodes 9 TB memory 125 Mbit/s (1 Gbit/s burst) Ethernet 4.5 PB of disk space, 4 PB tape 60K specints

the dutch e-science grid

BiG Grid





GPU Cluster (part of LISA)

Tesla GPU 2000 cores in 8 nodes 32 Gbyte memory (total for GPU) Infinband 20Gbit/s 2 Tbyte disk space 7 Tflop/s

HPC Cloud



i-R-O-D-S

hedoop

BiG Grid

Innovative Infrastructure

Hadoop CDMI Webdav, iRODS ClearSpeed



What is a HPC Cloud?



High Performance Computing Application Parallelization

Task & Data Parallelization



rid

the dutch e-science grid

Amdahl's Law: The speedup of a program using multiple processors in parallel computing is limited by the sequential fraction of the program.

For example, if 95% of the program can be parallelized, the theoretical maximum speedup using parallel computing would be 20× as shown in the diagram, no matter how many processors are used.

$$\overline{(1-P)+rac{P}{N}}$$

Processing Time



What is a Cloud?





What is Cloud Computing?

"Cloud computing is a model for enabling convenient, ondemand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction." (NIST-SP800-146)

Service Models:

- Cloud Software as a Service (SaaS)
- Cloud Platform as a Service (PaaS)
- Cloud Infrastructure as a Service (IaaS)

BiGGrid:

Allow users to

- _ freely instantiate a personal environment
- leap from laptop (small scale) to HPC (large scale)



What is a Cloud?

[National Institute for Standards and Technology NIST] [http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc]

- Resource Pooling,
 - Multiple concurrent users on a shared system
- Broad Network Access
 - Accesible from The Internet
- Measured Service
 - Pay per use
- Rapid Elasticity
 - Capabilities scaled up and down dynamically (pay-as-you-go)
- On-demand self-service
 - User is in full control



Is a Compute Centre a Cloud?

Yes

Yes

Yes

Some

NO

[National Institute for Standards and Technology NIST] [http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc]

• Resource Pooling,

Multiple concurrent users on a shared system

- Broad Network Access
 - Accesible from The Internet
- Measured Service
 - Pay per use
- Rapid Elasticity
 - Capabilities scaled up and down dynamically (pay-as-you-go) Within pre-allocation
- On-demand self-service
 - User is in full control No control over OS and adding resources not trivial



HPC Cloud vs HPC Cluster







BiG Grid the dutch e-science grid

Virtual Machines in a HPC Cloud Multi-tenancy Virtuele Machine Virtueel Cluster Virtuele Machine gebruiker C gebruiker A gebruiker B Fysiek Cluster Virtuele Machine Virtuele Machine Fysiek NODE Fysiek NODE





(HPC) Cloud Why?

World

- better utilization for infrastructure
- "Green IT" (power off under-utilization)
- easy management

BiGGrid

- free OS & software environment
- locked software can be used
- rapid availability
- HPC cloud for academic world

Massive interest and multiple early adopters prove the need for an academic HPC Cloud environment.

– beta-cloud running "production





HPC Cloud Philosophy

HPC Cloud Computing:

Self Service Dynamically Scalable Computing Facilities

Cloud computing is not about new technology, it is about new uses of technology







...At AMAZON?

- Cheap?
 - Quadruple Extra Large = 8cores and 64Gb ram: \$2.00/h (or \$5300/y + \$0.68/h)
 - -1024 cores = \$2.242.560/y (or \$678k + \$760k = \$1.4M/y)
- Bandwidth = pay extra
- Storage = pay extra
- I/O guarantees?
- Support?
- Secure (no analysis/forensics)?
- High Performance Computing??







What is needed to create a successful HPC Cloud?



Users of Scientific Computing

- High Energy Physics
- Atomic and molecular physics (DNA);
- Life sciences (cell biology);
- Human interaction (all human sciences from linguistics to even phobia studies)
- from the big bang;
- to astronomy;
- science of the solar system;
- earth (climate and geophysics);
- into life and biodiversity.





Slide courtesy of prof. F. Linde, Nikhef

Courrent) Users of HPC Cloud Computing

- High Energy Physics
- Atomic and molecular physics (DNA);
- Life sciences (cell biology);
- Human interaction (all human sciences from linguistics to even phobia studies)
- from the big bang;
- to astronomy;
- science of the solar system;
- earth (climate and geophysics);
- into life and biodiversity.





Slide courtesy of prof. F. Linde, Nikhef

HPC (Cloud) Application types

Туре	Examples	Requirements
Compute Intensive	Monte Carlo simulations and parameter optimizations, etc	CPU Cycles
Data intensive	Signal/Image processing in Astronomy, Remote Sensing, Medical Imaging, DNA matching, Pattern matching, etc	I/O to data (SAN File Servers)
Communication intensive	Particle Physics, MPI, etc	Fast interconnect network
Memory intensive	DNA assembly, etc	Large (Shared) RAM
Continuous services	Databases, webservers, webservices	Dynamically scalable





• Develop together with users



The product: Virtual Private HPC Cluster

• We (plan to) offer:

- Fully configurable HPC Cluster (a cluster from scratch)
- Fast CPU
- Large Memory (64GB/8 cores)
- High Bandwidth (10Gbit/s)
- Large and fast storage (400Tbyte)
- Users will be **root** inside their own cluster
- Free choice of OS, etc
- And/Or use existing VMs: Examples, Templates, Clones of Laptop, Downloaded VMs, etc
- **Public** IP possible (subject to security scan)

BiG Grid the dutch e-science grid



Platform and tools:

- Redmine collaboration portal
- Custom GUI (Open Source)
- Open Nebula + custom add-ons
- CDMI storage interface











Virtual architecture User view



HPC Cloud trust (1/2)

Security is of major importance

- cloud user confidence
- infrastructure provider confidence

Protect

- the outside from the cloud users
- the cloud users from the outside
- the cloud users from each other
- Not possible to protect the cloud user from himself
 - user has full access/control/responsibility
 - ex. virus research must be possible



HPC Cloud trust (2/2)

- Firewall
 - fine-grained access rules ("closed port" policy)
 - non-standard ports open on request only and between limited network ranges
- Scanning of new virtual templates
 - catches initial problems, but once the VM is live...
- Port scanning
 - catches well-known problems
- State-full Package Inspection
 - random sample based



Project Development Goals

- Physical Architecture
 - HPC Cloud needs High I/O capabilities
 - Performance tuning: optimize hard- & software
 - Scheduling
- Usability
 - Interfaces
 - Templates
 - Documentation & Education
 - Involve users in pre-production (!)
- Security
 - Protect user against self, fellow users, the world and vice versa!
 - Enable user to share private data and templates
 - Self Service Interface
 - User specifies "normal network traffic", ACLs & Firewall rules
 - Monitoring, Monitoring, Monitoring!
 - No control over contents of VM
 - monitor its ports, network and communication patterns

BiG Grid

the dutch e-science grid

ROADMAP

- 1) SARA Innovation project in 2009,
- 2) Pre-production for BiGGrid in 2010
- 3) In 2011 (summer) Production Infrastructure
- 4) Development continues 2011/2012

A Team and a bit of Hard Labour











User collaboration Portal

Redmine (www.redmine.org)



the dutch e-science grid

Self Service GUI

	Applications	Places System	300	30						C 4 0 92'C C	The Dec	1, 19:18 = jerces
	Statement Statement			CHARLES	Open	Nebula Manapo	ment Co	roole - Mozilla Firefor				
	tide Est Vice											
	🔶 🕂 🕂 d		Berari	https://ul.claud	basaran(hosts.ph					+ - 🏭 Goog		, a
Applications Places System 3 Ball	NUN Algemee	n Ni 🖪 Sieshe	lot 📑 Knare	ce 🗎 Subtrac H	IPCV Canglus !	зала нис 🗎 О	idtip 🛢	SAM: Data transfers 🗎	UHC Status 🗎 Va	wgerdinen op m 🗎 1	Later Bick Assure .	- A.
OpenNebula Management Console - Mocilia Firefox	OpenNebul	a Manager	nent C	onsole								admin - logout
niffe Edt Vew Highly Booknads Bolt Brip	vm's hosts r											100
🚸 🗄 - 🗘 🔿 🌦 🔳 Sector Algocht deuts secuti	Cloud hosts:								Create host:			
💈 NUH Algemeen NL 📲 Stashdot 🛢 Inance 🛢 Subtrac HICH 🗑 Conglia: SARA HIC 🛢 Driditap 🖥 SAM- Data transfers 🗎 UHC Status 🖥 Youwgordje	len our Id Name Runnie	ng VM Total op	u Free cp	u Assigned i	pu Total memo	ry Free memo	ry Status	5	Hostname:			
OpenNebula Management Console	0 host01 1	800	800	800	3262916	3012120	on	show disable - ok	IM:	im_kom ~		
vm's hosts networks users	1 host02 0	800	800	800	3262916	3254028	on	show disable - ok	VHM:	amm_kam ~.		
Cleud vm's: Create vm:	2 host03.0	800	800	800	3262916	3206799	on	show disable ~ ak	TM:	tm_ush ~		
Id User Name Status Cpu Memory Host Days Time	3 host04.6	Openblah		304	nt Consol	3451053		start franks and all	a contractor			
194 cloud12 ubuntu-9 shut 0 2097152 host05 22 05:11:30 show delete v none v 04	4 host058							Togge		gont pression 1.0.1		
515 cloud14 ubuntu-9 runn 0 2097152 host06 07 06:19:16 show suspend v none v ok	5 host06.3	Upload image		figuration	disk image uploa	ad disk image	man ay a	ament hosts nëtw	ronis quotas			
701 oneadminidebian runn 0 2097152 host05.06 01:38:59 show suspend + rone + ok		upranu rinige.	34									
740 onesidminidebian runn 0 2097152 host05 05 05:15:21 show suspend v none v ok		1 Jum	pLoade	r								
743 cloud17 hadoog-1 runn 0 2097152 host05 05 04:26:33 show septend v none v ek	ions Places System		-							K	1, 19:17 = jerces	
744 cloud17 hadoop-2 runn 0 2097152 host05 05 04:26:33 show suspend v none v ck	A STRATE RESIDENCE AND A STRATE RESIDENCE	Paste	0	Adc	🖸 Flemov 🔒	Retry fa 😋	91	10 20	<u>a o</u>		(a)(a)(a)	
745 cloud17 hadoop-3 runn 0 2097152 host04 05 04:26:33 show suspend v rose v ck	View History Bookma	(S) 1	3		× -			Console	- Mozilia Elrefox		<u></u>	
771 cloud09 ubuntu-9 stop 0 2097152 host0104 04:17:12 show restart v rene v sk	· • • 👘 🔳				1 5 se	ra.nl https://ui.clat	idia.sara.n	l/console.php?id=181			<u>م</u> [ش	
774 cloud09 ubuntu-9 stop 0 2097152 host04 04 03:45:13 show restart v nore v 64	gemeen Ni 🖪 Sieshdot		1		Disc of Foreign	rne Octors (ly from a kil lina	poored D	acono - Dano (14-41-45al B 18	я і я-І			
781 cloud09 ubuntu-9 stop 0 2097152 host04 04 02:56:21 show restart + none + at OpenNo	ebula Manageme				/deu/	sda5: clean, 1 Fran utlJ-linu	64/12449	96 Jiles, 20010/2409	076 blocks		pout	
	ts networks users				zili sez	napper/elundur Lling prolimin	die ennet	: clcan, 16008/60960	00 filox, 2260	47/2436896 Minuks		
1045 cloud09 ubuntu-9 runn 0 2097152 host01 00 03:59:23 show suspend v rene v ek Cloud netw	orksi	dəb ar50	File Fr	dit View His	tory Bookmarks			1.2				
1055 cloud12 hadoop-m runn 0 2097152 host06 00 03:42:32 show selpend v neee v ek Id User	Name Type Br	1.22		enNebula Mar	agement Cons	4				τ		
1056 cloud12 hadoop-s runn 0 2097152 host05 00 03:41:31 show suspend + none + ek 0 oneadm	nin public Flored br		Opent	Nebula Mani	agement Conso	le						
	in private_01 Fixed br				uselan – disk imago upi	nad - risk image man	ngemere d	Logged in a hostris norwarks public fore	u dauda - loquit vario will exceptions quar	n: 1.0.1 ac		
1058 doud12 hadoop-s runn 0 2097152 host05 00 03:41:31 show asspend v nees v ek 9 oneadm 1059 doud12 hadoop-s runn 0 2097152 host04 00 03:41:31 show asspend v nees v ek	in private_02 Fixed br	(Δ)	Creater()	Notity template:								
100 chedril 2 hadoop-s runn 0 2097152 host04:00 03:41:31 show supped v name v at	in private_03 Fixed br						e a tempta	ie by changing the name di	ring etiling			
11 oneadm	nin privabe_04 Flood br	3	Rene:			totonal_VM_public						
🗶 🔄 Hibos (1941) total) - Ex. 🔰 Re: Opennebula Mana. 😽 OpenNebula Manage. 🔠 jeroengitus -	nin private_05 Fixed br	Goto uploa S*	Cou:			1 7						
			OS us his	381:								
			' Boot:			hd y					-	
Developed at SARA			Disk sect									
			Number Network:			1 🔻						
Open Source, available at				of networks:		2 1						
			Graphics	section:		2 La						
www.opennebula.org			VNC Pas	swend:		1 10						
epennebulatorg			neve								24	



Monitoring workload





Gangla Web Frontend version 3.1.7 Check for Updates. Gangla Web Bickend (gmetzd) version 3.1.7 Check for Updates. Downloading and parting gangla's XML the book 0.0288. Images ceated with REDTool version 1.3.1. Pages generated using TemplatePower version 3.0.1.

Standards: OCCI + CDMI + OVF + CNMI = CMI


Development plans/effort @ SARA

- Storage
 - CDMI server application
- Network
 - Dynamic provisioning
 - QoS
 - ACL/Firewall rules
 - Dynamic DNS
 - "CNMI"
 - Network benchmarking
- Compute
 - OCCI server with AAA?



- GUI
 - New & improved on OCCI/CDMI
- Security
 - Flow analysis
 - Dynamic ACL/Firewall

Real world network virtualization tests with qemu/KVM

- 20 gbit/s DDR infiniband (IPoIB) is compared with 1 Gbps
 Ethernet and 10 Gbps Ethernet
- Virtual network bridged to physical (needed for user separation)
- "real-world" tests performed on non optimized system
- Results
 - 1GE: 0,92 Gbps (1 Gbs)
 - IpoIB: 2,44 Gbps(20Gbs)
 - 10GE: 2,40 Gbps (10Gbs)
- Bottleneck: virtio driver
- Likely Solution: SRIOV
- Full report on www.cloud.sara.nl



User participation 30 involved in testing

nr.	Title	Core Hours	Storage	Objective	Group/instiute		
1	Cloud computing for sequence assembly	14 samples * 2 vms * 2-4 cores * 2 days = 5000	10-100GB / VM	Run a set of prepared vm's for different and specific sequence assembly tasks	Bacterial Genomics, CMBI Nijmegen		
2	Cloud computing for a multi- method perspective study of construction of (cyber)space and place	2000 (+)	75-100GB	Analyse 20 million Flickr Geocoded data points	Uva, GPIO institute		
3	Urban Flood Simulation	Field		# projects	al Science		
4	A user friendly cloud-based inverse modelling environment Real life HPC cloud computing	Bioinformatic	S	8	o-ecology,		
5	experiences for MicroArray analyses	Ecology		3	matics Únit,		
6	Customized pipelines for the processing of MRI brain data Cloud computing for historical	Geography		ng Group, mus MC			
7	map collections: access and georeferencing Parallellization of MT3DMS for	Computer sc	ience	5	raphy, UvA		
8	modeling contaminant transport at large scale						
9	An imputation pipeline on Grid Gain	Linguistic	CS	4	formatics f groningen		
10	Regional Atmospheric Soaring Prediction	Other	o-ecology,				
11	Extraction of Social Signals from video	160	630GB	Video Feature extraction	Laboratory, TU Delft		
12	sequencing data from mouse tumors	?	150-300GB	Run analysis pipeline to create mouse model for genome analysis	Chris Klijn, NKI		



Example Project 1

- Medical data MRI Image processing pipeline
 - Cluster with custom imaging software
 - Dynamic scaling up depending on the load
 - Added 1 VM with web service for user access, data upload and download

BiG Grid

s from H. Vrene dutter Science grid





Example project 2 NMR spectroscopy: Virtual Cing by J. Doreleijers

With NMR spectroscopy the 3D structure of biomolecules such as proteins and DNA are solved in solution. It thus provides a structural view of the chemical reactions that underly most diseases.

NMR structure determination needs a solid validation of the experimental data in relation to the resulting 3D coordinates because the process in many labs has not and often -can- not be automated fully. A virtual machine called VirtualCing (VC for short) interfaces to the best 24 NMR validation programs, together with CING's internal unique checks. VC was developed because installing the external programs on a traditional grid would take too long in development and would be cumbersome to maintain. We were able to validate all the 8,000+ structures currently available in the worldwide database Protein Data Bank (wwPDB) in just a week. The same strategy is applied to recalculate, improve and validate several thousand protein structures in a new project named NMR_REDO.

1AON		1011	mat	1/13	R	1A1D		146D		148D		149D	Z	156D	
LAIP	2	IAIT		IAIU	- Constant	IAIW	See B	169D	X	170D		171D		175D	X
LAIZ	Sec.	1A23		1424		1421 1291		L76D	E	1790	6	17RA	Ł	1850	T
1A2S	Sur.	1A3M	No.	1A0P 4201	- Contraction of the second se	1A4D	No.	L86D	×	1930		1990	Ş	1A03 4430	R



User Experience

(slides from Han Rauwerda, transcriptomics UVA)

Microarray analysis: Calculation of F-values in a 36 * 135 k transcriptomics study using of 5000 permutations on 16 cores. Over 10 week period **30.000** core-hours Data analysis using R (statistical analysis) with specialized plugin



Ageing study - conditional correlation

- dr. Martijs Jonker (MAD/IBU), prof. van Steeg (RIVM), prof. dr. v.d. Horst en prof.dr. Hoeymakers (EMC)
- 6 timepoints, 4 tissues, 3 replicates and 35 k measurements + pathological data
- Question: find per-gene correlation with pathological data (staining)
- Spearman Correlation conditional on chronological age (not normal)
- p-values through 10k permutations (4000 core hours / tissue)

Co-expression network analysis

- 6k * 6k correlation matrix (conditional on chronological age)
- calculation of this matrix parallellized. (5.000 core hours / tissue)

Development during testing period (real life!)

Conclusions

- Many ideas were tried (clusters with 32 64 cores)
- worked out of the box (including the standard cluster logic)
- no indication of large overhead
- Cloud cluster: like a real cluster
- Virtually no hick-ups of the system, no waiting times
- User: it is a very convenient system



the dutch e-science grid

Our of the statistics in beta phase

Users liked it:

- ~90.000 core-hours used in first 10 weeks (~175.000 available)
- currently 500k core-hours
- 50% occupation during beta testing
- Currently 80-90%
- Science is being done!
- Some pioneers paved the way for the rest ("Google" launch approach)



Observations

- Usage: Scientific programmer prepares environment, Scientist uses
- Several "heterogenic clusters" Microsoft Instances combined with Linux
- Modest parallelism (maximum 64)
- User wishlist: Possibility to share a collection of custom made virtual machines with other users
- Added value: support by your trusted HPC centre.
- HPC Cloud on HPC hardware is necessary addition to a complete HPC eco-system



Advantages of HPC Cloud

- Only small overhead from virtualization (5%)
- easy/no porting of applications
- Applications with different requirements can coexist on the same physical host
- Long running services (for example databases)
- Tailored Computing
- Service Cost shifts from manpower to infrastructure
- Usage cost in HPC stays Pay per Use
- Time to solution shortens for many users

BiG Grid the dutch e-science grid

HPC Cloud, what is it good for?

- Interactive applications
- High Memory, Large data
- Same data, many different applications (Cloud reduces porting efforts!)
- Dynamic, fast changing and complicated applications
- Clusters with Multi Operating Systems
- Collaboration
- Flexible and Versatile
- System architecture is expandable and scalable

BiG Grid the dutch e-science grid

BiG Grid HPC cloud in international media



the dutch e-science grid





SNEAK PREVIEW (What is an ideal system for an HPC Cloud)



Calligo "I make clouds"

19 Nodes:

- _ CPU Intel 2.13 GHz 32 cores (Xeon-E7 "Westmere-EX")
- _ RAM 256 Gbyte
- _ "Local disk" 10 Tbyte
- _ Ethernet 4*10GE



Total System

- _ 608 cores
- _ RAM 4,75TB
- 96 ports 10GE, 1-hop, nonblocking interconnect
- 400TB shared storage (ISCSI,NFS,CIFS,CDMI...)
- _ 11.5K specints / 5TFlops





BiG Grid the dutch e-science grid

Calligo, system architecture



the dutch e-science grid





HPC Systems @ SARA

System	Node	Total
Huygens National Super	CPU Power6, 4.7Ghz, 32/64 cores RAM 128/256 GB "local disk" 8Tbyte Infiniband 8*20 Gbit/s	3456 cores in 105 nodes 15.25 TB of memory, 700 TB of disk space, 60 TFlop/s
LISA National Compute Cluster	CPU Intel 2.26Ghz 8/12 cores RAM 24 Gbyte Local disk 65Gbyte/200Gbyte Infiniband 20Gbit/s	4480 cores in 512 nodes, 12 TB of memory, Infiniband 20Gbit/s 50Tbyte disk space 20 TFlop/s
Grid	CPU Intel 2.2 Ghz 8 cores RAM 24 Gbyte Local disk 300Gbyte Ethernet 1 Gbit/s	3000 Cores in 3000 nodes 9 TB memory 125 Mbit/s (1 Gbit/s burst) Ethernet 4.5 PB of disk space, 4 PB tape 60K specints
Cloud	CPU Intel 2.13 GHz 32 cores (Xeon-E7 "Westmere-EX") RAM 256 Gbyte "Local disk" 10 Tbyte Ethernet 4*10GE	608 cores RAM 4,75TB 400TB shared storage (ISCSI,NFS,CIFS,CDMI) 11.5K specints / 5TFlops 96 ports 10GE, 1-hop, non- blocking interconnect



Acknowledgements



Our Sponsor: NL-BiGGrid Our Brave & Entrepeneurial **Beta Users** And the HPC Cloud team: Jhon Masschelein, Tom Visser, Dennis Blommesteijn, Neil Mooney, Jeroen Nijhof, Alain van Hoof, Floris Sluiter et. al.

http://www.cloud.sara.nl





Demo





SIMPLY EXPLAINED - PART 17: CLOUD COMPLITING

BiG Grid the dutch e-science grid

photo: http://cloudappreciationsociety.org/