Ph.D. Qualifying Exam

Hyungro Lee

School of Informatics and Computing

Indiana University - Bloomington

Topics

- Virtualization
- Monitoring Distributed Systems
- Bioinformatics Applications in The Cloud

Virtualization

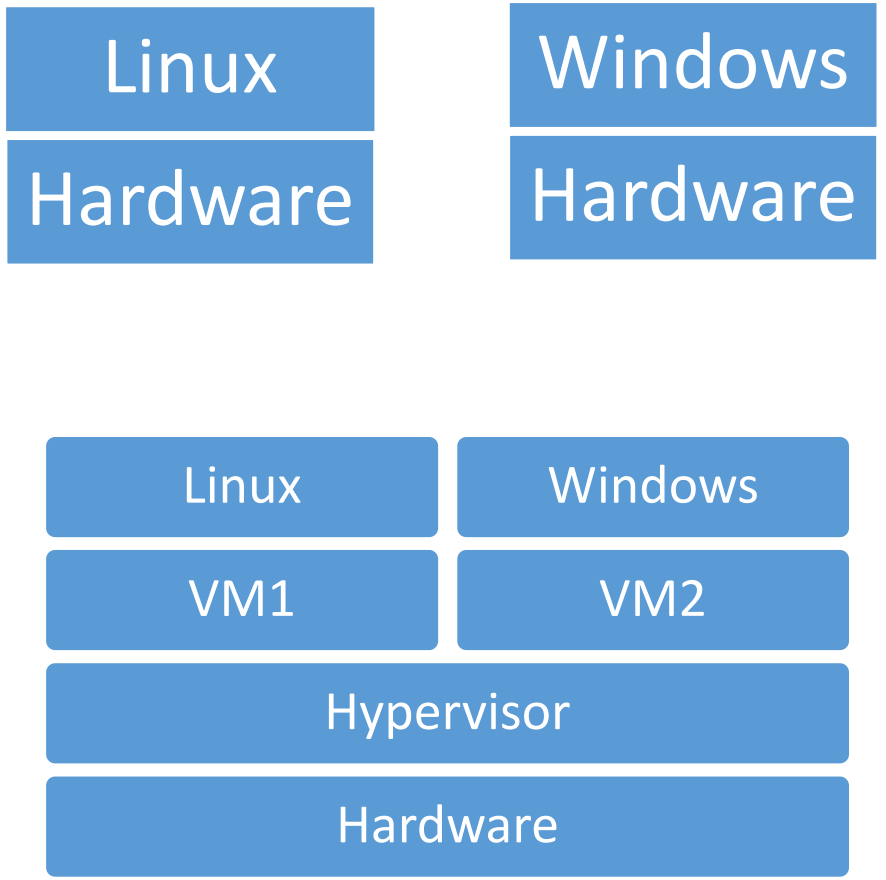
Hypervisor and Resource Virtualization

Virtualization

- Definition
- Terminology
- Type of virtual machine monitors (hypervisors)
- CPU virtualization
- Memory virtualization
- I/O virtualization

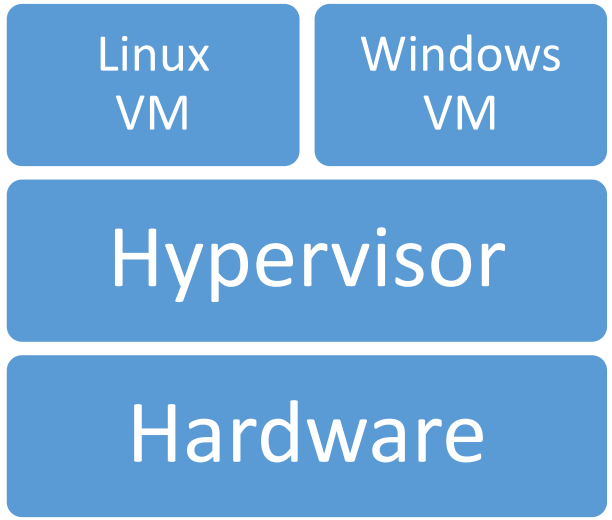
Virtualization is

- a process of creating virtualized environment with additional interface so that operating systems can run with the virtual resources separating from physical hardware.
- In virtualization, the additional interface is **hypervisor** (also called virtual machine monitor) which creates a **virtual machine**(s).



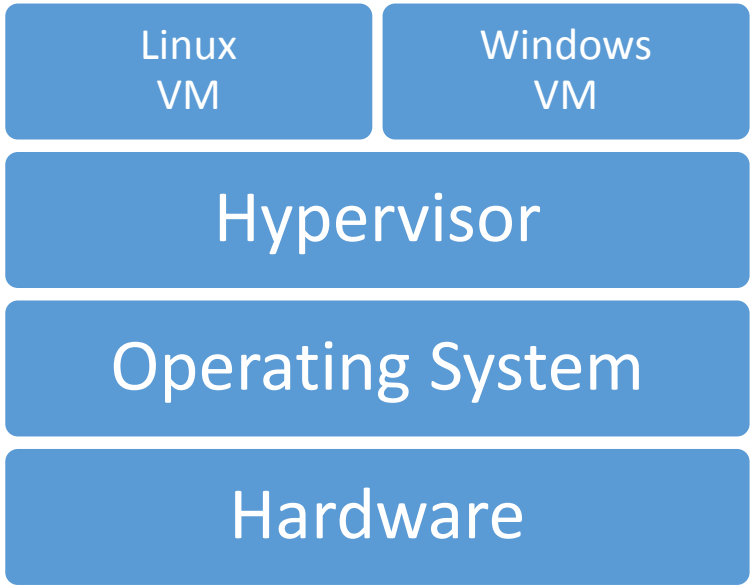
Hypervisor (VMM)

- Type 1 Hypervisor (Bare metal)
  - Tied with physical hardware without operating system
  - Install virtual instance using management console
  - VMWare vSphere, Microsoft Hyper-v



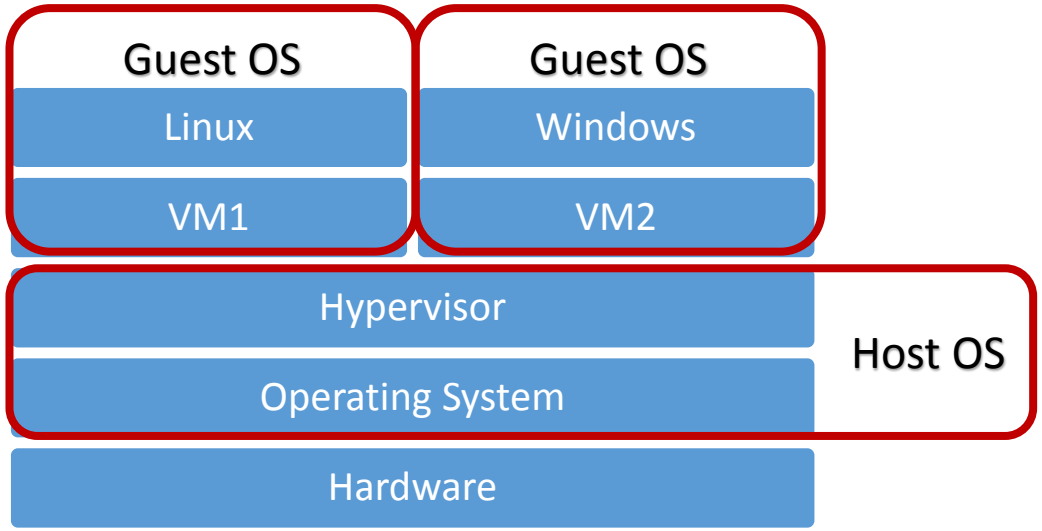
Hypervisor (VMM)

- Type 2 Hypervisor (Hosted Hypervisor)
  - Tied with a physical machine on top of operating system
  - Hypervisor runs on the operating system like other normal software e.g. Microsoft Word, Adobe Acrobat PDF
  - VirtualBox, VirtualPC, VMware Workstation



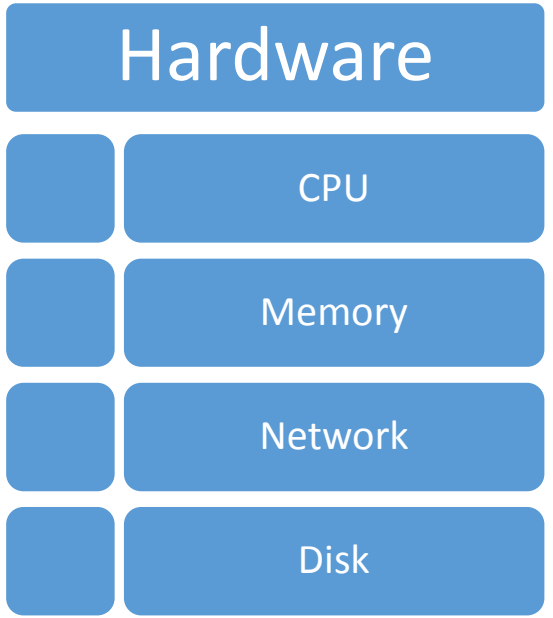
Terminology

- Guest OS/ Host OS
- Hypervisor (also Virtual Machine Monitor, VMM)
- Virtual Machine (VM)
- Full virtualization
- paravirtualization
- Hardware-assisted (HVM)
- Unmodified guest
- Xen, VMware offers virtualization



Resource virtualization

- CPU
- Memory
- Network (I/O devices)
- Disk

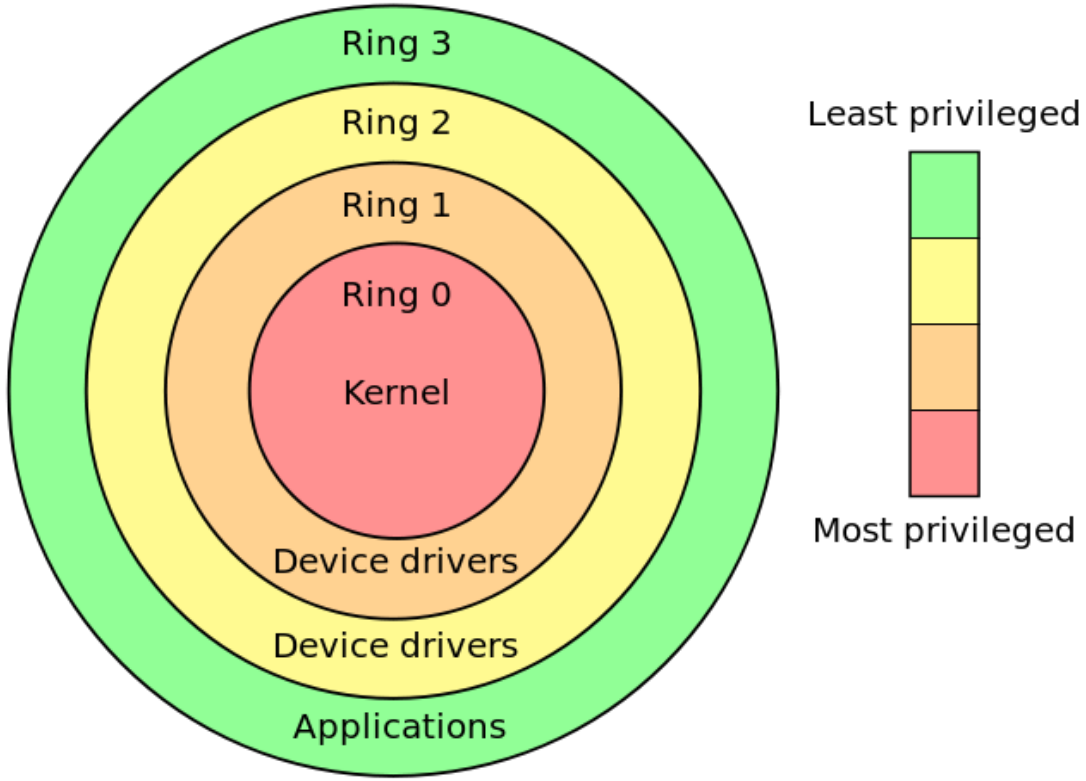


CPU Virtualization

- X86 architecture
- Full virtualization
- Paravirtualization
- Hardware-assisted virtualization
- Ring privilege level – assembly instruction set

Ring privilege level

- x86 architecture
- Memory, CPU, I/O ports protected
- Operating system
  - Kernel code runs in ring 0
- User application
  - Runs in ring 3
- Most modern x86 kernels use only two levels ring 0 and 3

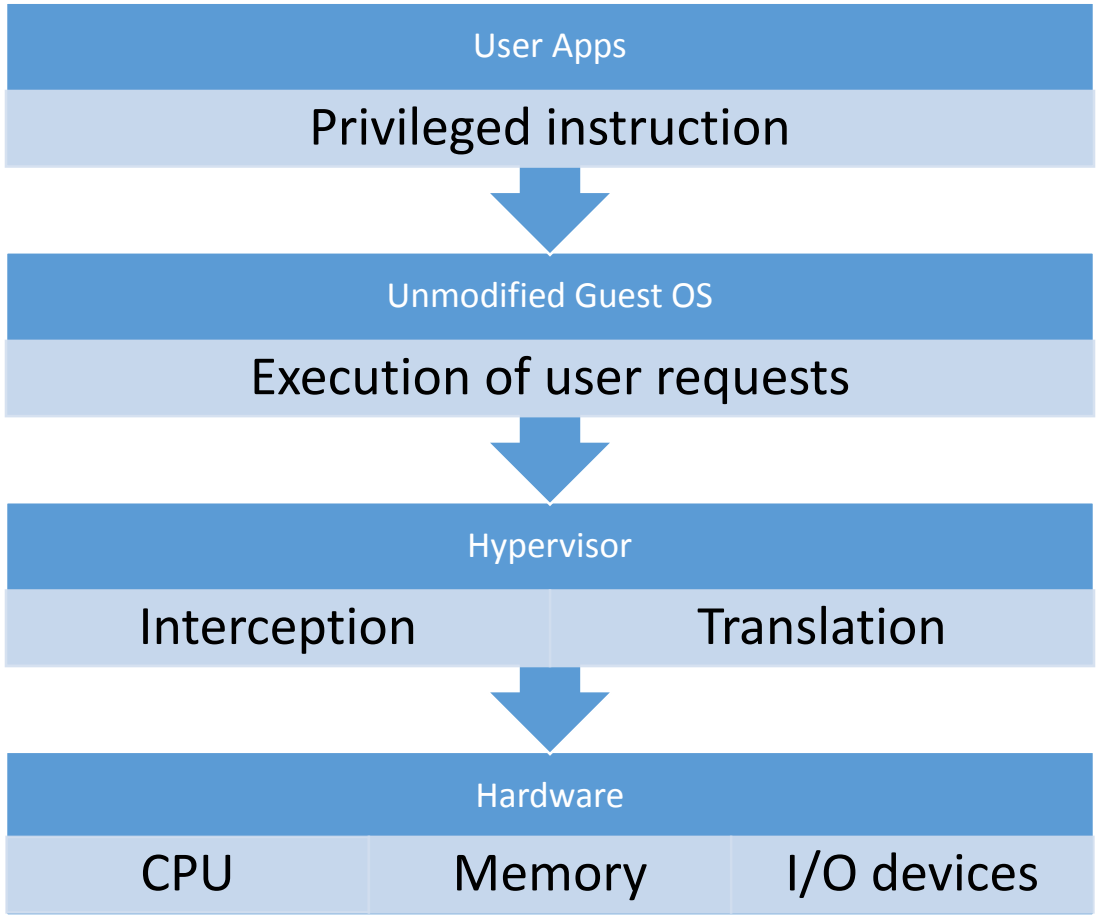


Privilege rings for the x86 available in protected mode

Image source: Wikipedia.org

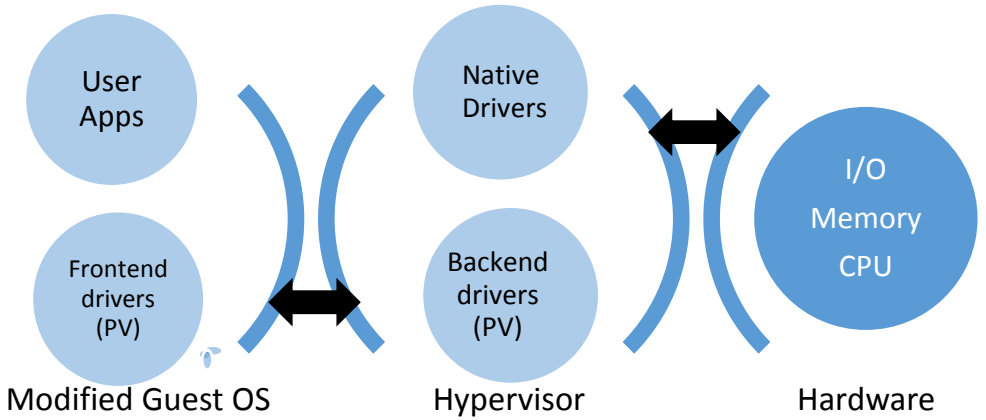
Full Virtualization

- Unmodified guest OS
- Privileged instructions intercepted and translated by VMM
- Binary translation by VMware
  - (guest OS is in ring 1, VMM is in 0)



Paravirtualization

- Modified guest OS
  - Kernel and driver
- Better performance but compatibility issue
- VMWare VMI, Xen PVOPS

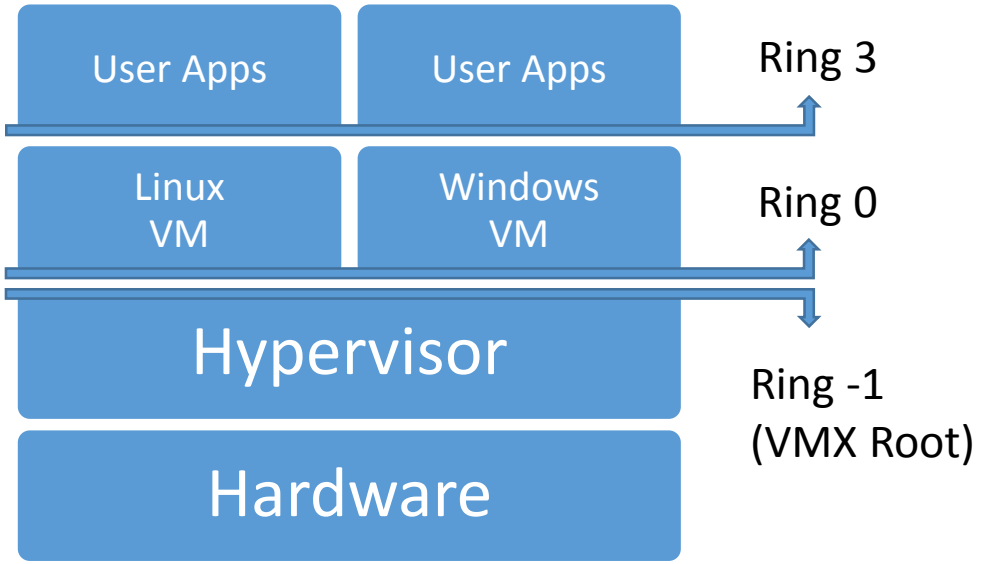


Xen's para-virtualization with control domain (dom0)

Barham, Paul, et al. "Xen and the art of virtualization." *ACM SIGOPS Operating Systems Review* 37.5 (2003): 164-177

Hardware-assisted Virtualization

- Intel VT-x, AMD-v
- No kernel modification
- No binary translation
- Virtual Machine eXtension (VMX)
- New privilege level beneath Ring 0



CPU Supports for Hardware Acceleration

- INTEL VTX -> vmx
- AMD-V -> svm

```
$ egrep '(vmx|svm)' /proc/cpuinfo
flags		: fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts
acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx lm constant_tsc arch_perfmon pebs bts rep_good
aperfmperf pni dtes64 monitor ds_cpl vmx smx est tm2 ssse3 cx16 xtpr pdcm sse4_1 xsave lahf_lm
tpr_shadow vnmi flexpriority
```

Check virtualization support in Linux

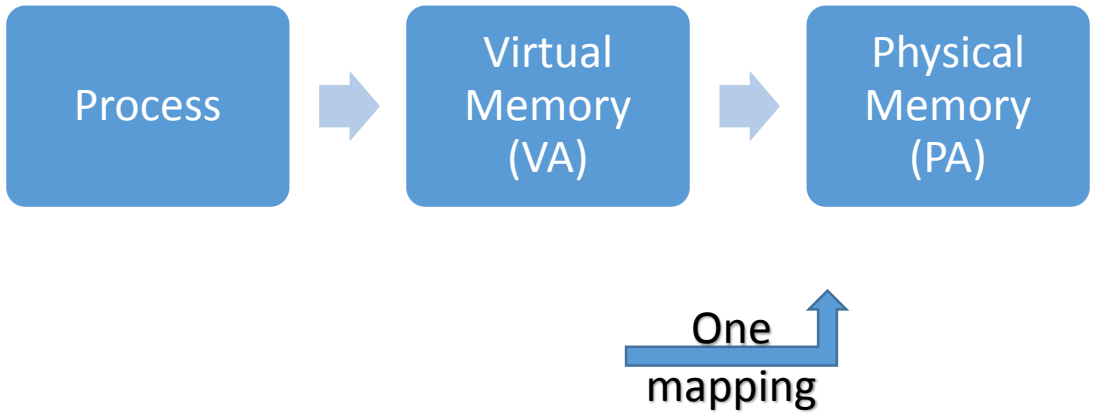
Different Approaches to Virtualization

	Full Virtualization	Para-Virtualization	Hardware-assisted Virtualization
Technique	Trap and translation for privileged instructions e.g. binary translation by VMware	Kernel and driver modification in OS e.g. pv driver, hypercalls	VM Exit to VMX Root mode on privileged instructions e.g. Intel VT-x, AMD-v
Implementation	VMware Workstation, Win4Lin Pro	Xen, VMware	VMware, Xen, Microsoft, Parallels
Guest OS	Unmodified runs in ring 1	Modified runs in ring 0	Unmodified runs in ring 0
VMM	Ring 0	Below Ring 0	Ring -1
Compatibility		With kernel and driver support (OS modification)	Hardware acceleration (vmx, svm)

Horne, Chris. "Understanding Full Virtualization, Paravirtualization, and Hardware Assist." *White paper, VMware Inc* (2007).

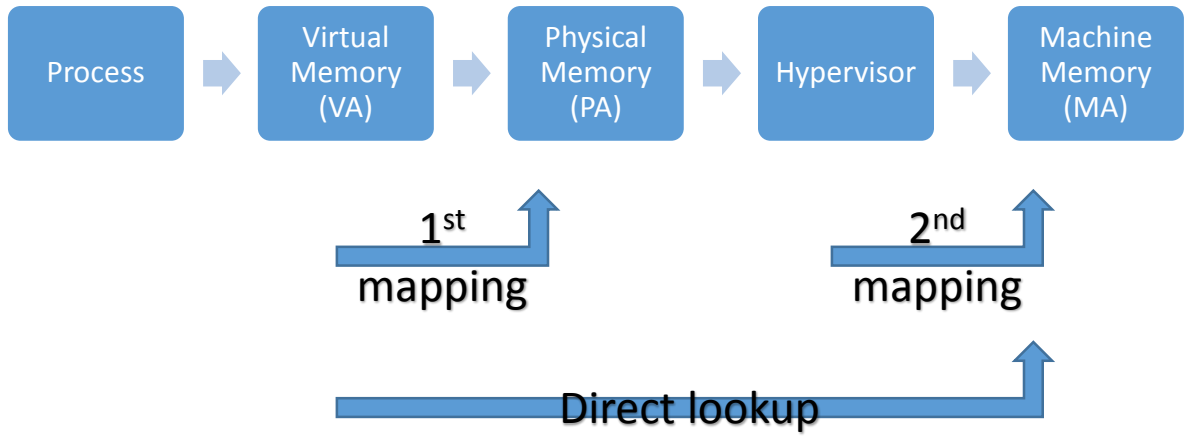
Memory Virtualization

- In a traditional execution, 1-stage mapping
- Memory Management Unit (MMU) stores a cache
- Translation Lookaside Buffer (TLB) is the cache
- Page Table stores mapping information



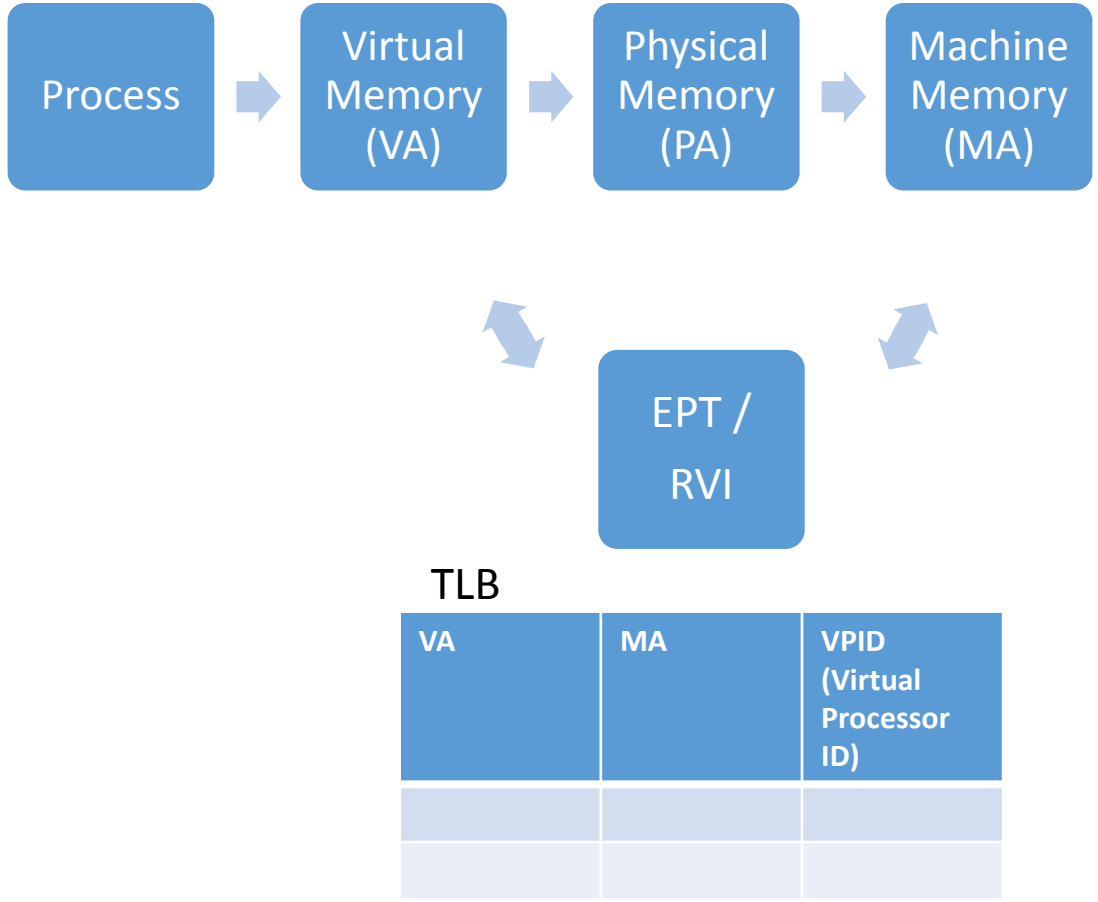
Memory Virtualization

- 2-stage memory mapping
  - Overhead
- 2nd page table to store the mapping between PA and MA in hypervisor
- Shadow Table used to store direct mapping from VA to MA (or HA)



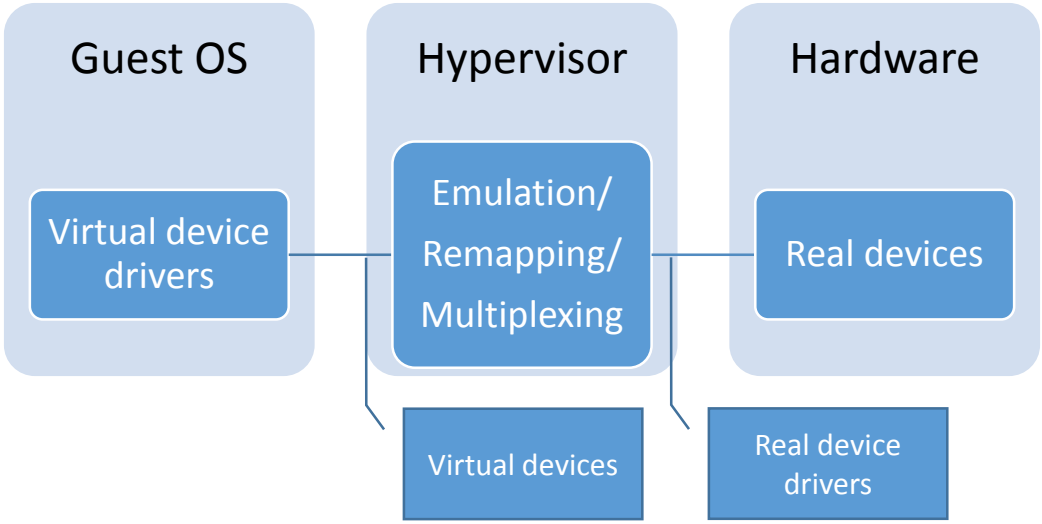
Memory Virtualization (Hardware-assisted)

- AMD's RVI (Rapid Virtualization Indexing)
- Intel's EPT (Extended Page Table)



I/O Virtualization

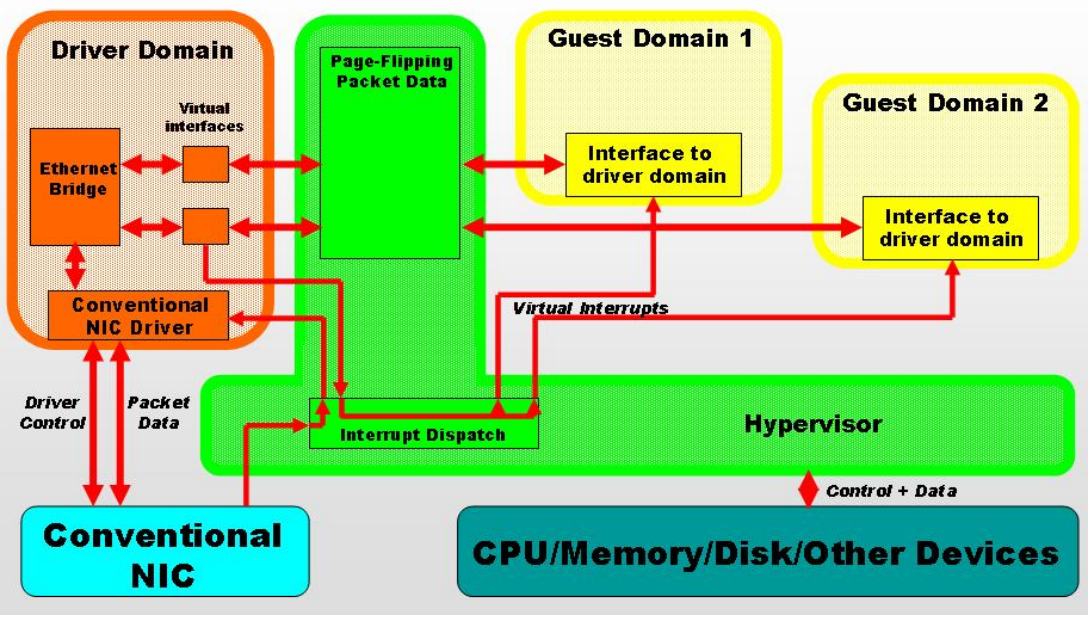
- Device emulation
- Mapping I/O addresses
- Multiplexing



V. Chadha, R. Illikkal, R. Iyer, I/O Processing in a virtualized platform: a simulation-driven approach, in: Proceedings of the 3rd International Conference on Virtual Execution Environments (VEE), 2007
Y. Dong, J. Dai, et al., Towards high-quality I/O virtualization, in: Proceedings of SYSTOR 2009, The Israeli Experimental Systems Conference, 2009.

I/O Virtualization (Xen)

- Control Domain (Domain 0)
- Domain U for Guest OSes (VMs)

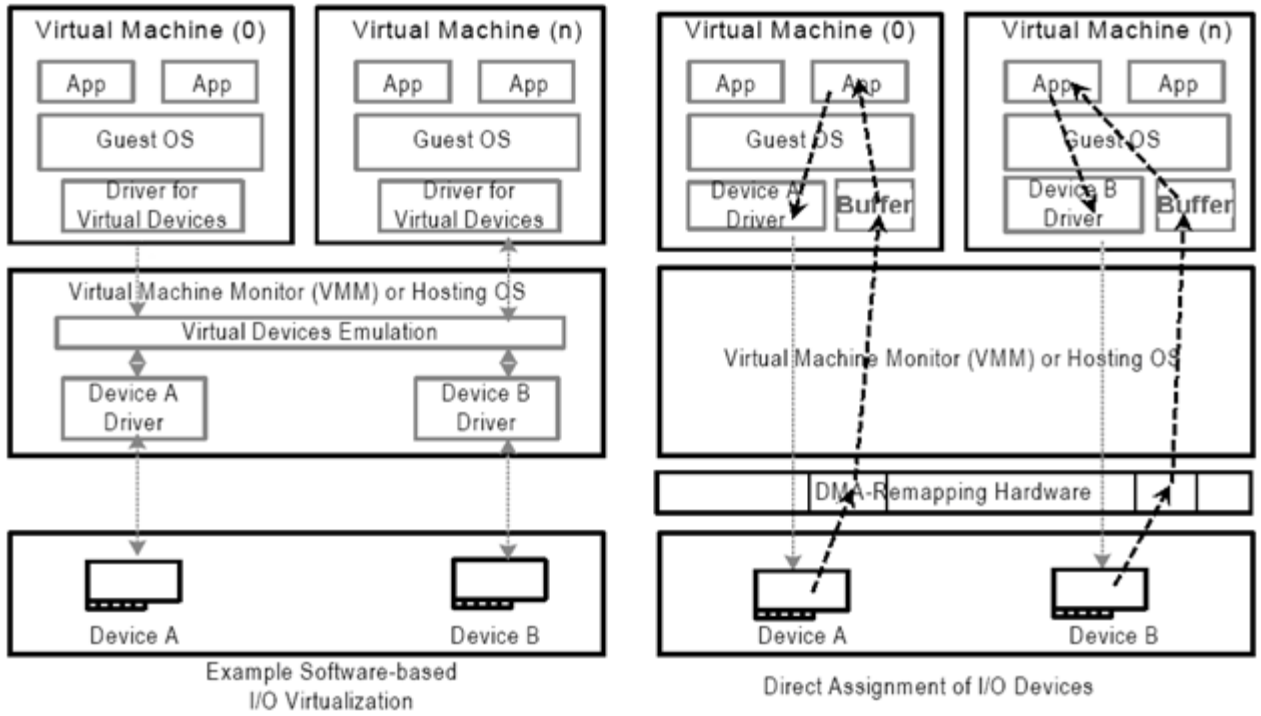


Xen virtual machine environment

Image source: Willmann, Paul, et al. "Concurrent direct network access for virtual machine monitors." *High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on*. IEEE, 2007.

I/O Virtualization (Hardware-assisted)

- I/O MMU for DMA Address Translation and protection
- Intel VT-d, VT-c (Virtualization Technology for Directed I/O and for Connectivity)
- AMD I/O MMU

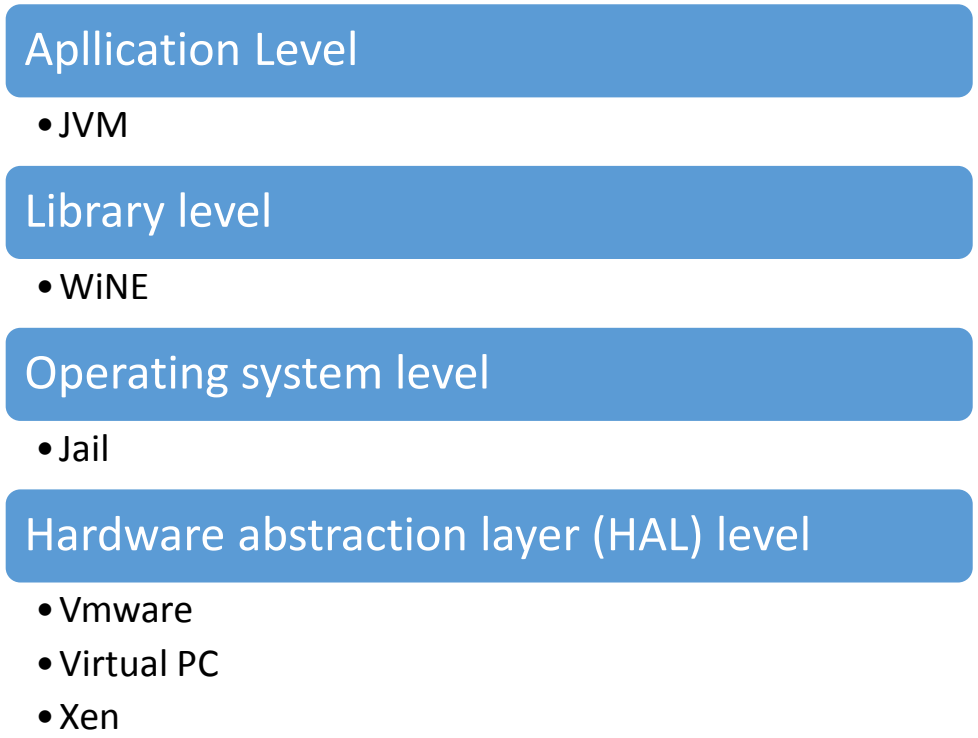


Software Emulation based I/O vs. Hardware based Direct Assignment I/O

Image source: Intel® Virtualization Technology for Directed I/O (VT-d): Enhancing Intel platforms for efficient virtualization of I/O devices by TW Burger

Level of Virtualization

- Different level of virtualizations
  - Application level (jvm)
  - Library level (WINE)
  - OS-level
- In the cloud, virtualization means **server virtualization**
  - Server consolidation



Summary of Virtualization

- Isolation from hardware
- Key component of cloud computing (but not identical)
- Software, hardware, or hybrid implementation for virtual resource management

Monitoring Distributed Systems

Grid, Clusters and Cloud

Monitoring Distributed Systems

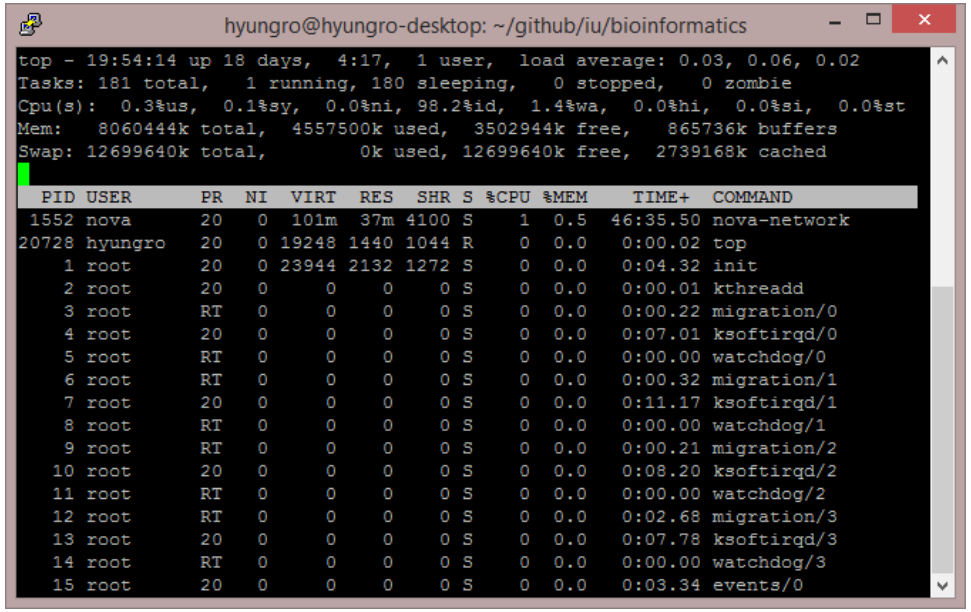
- Background
  - Definition
  - Design challenges
- Architecture
  - Implementation
- Monitoring in the cloud
  - Examples

Monitoring is

- A process to collect performance data and resource usage
- Detect problems
- Notification
- Estimate capacity planning

PC Monitoring

- Standalone PC
  - Mac User Activity Monitor
  - Windows Performance Monitor



Linux top command

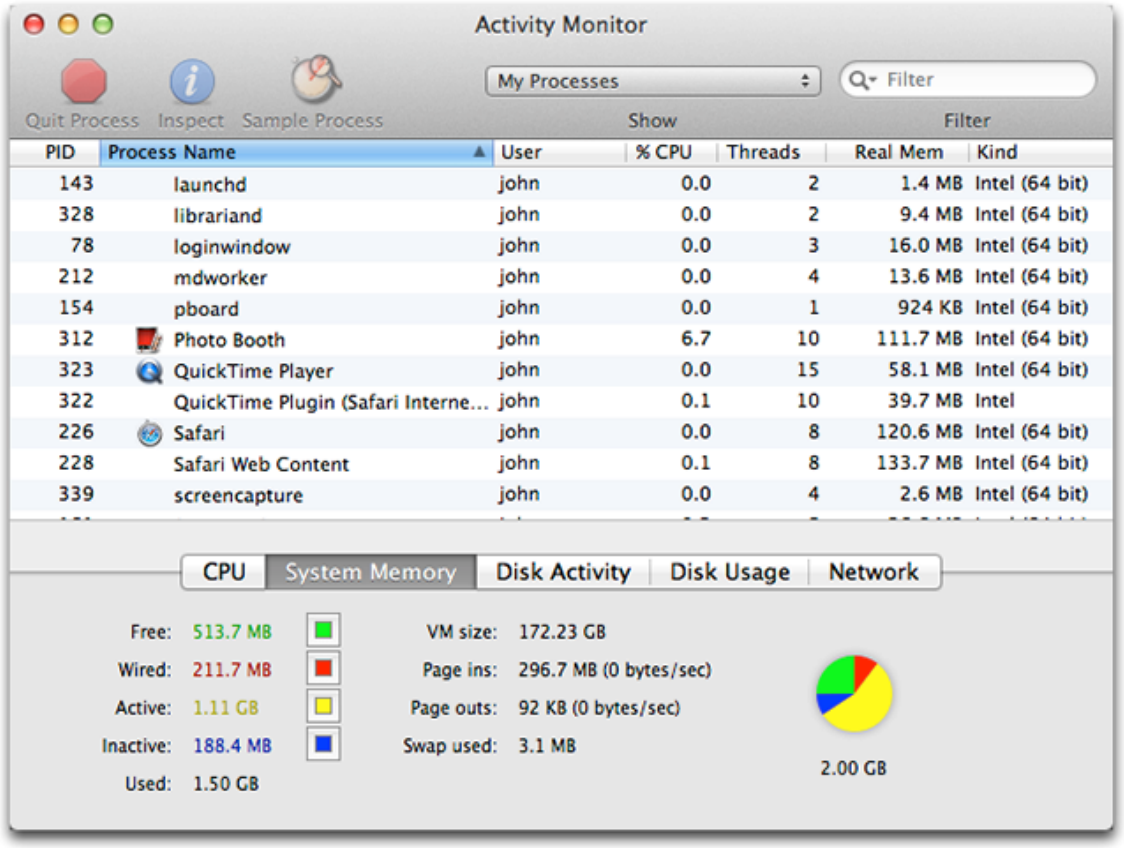
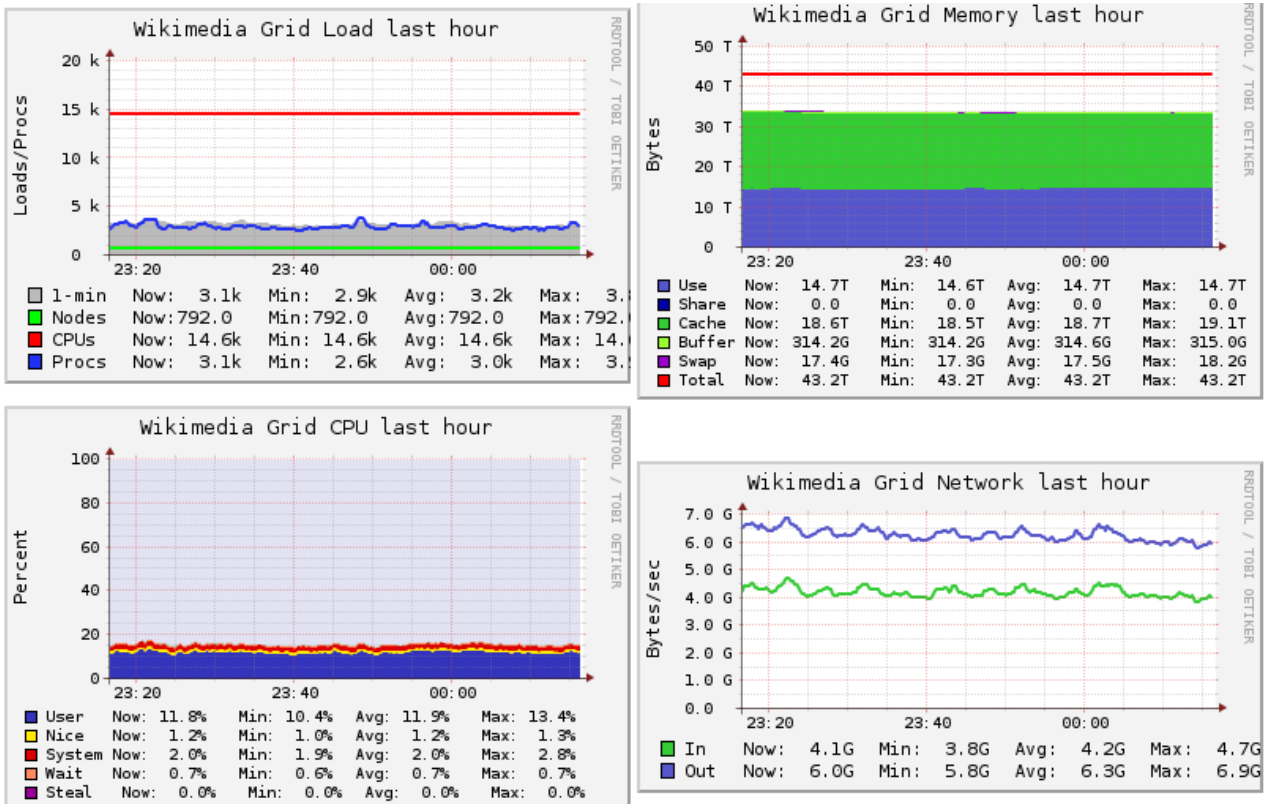


Image source: apple.com

Grid Monitoring

- Grids and clusters
  - Ganglia
  - Nagios
- Low overhead, latency



Load, Memory, CPU, Network monitoring by Ganglia

Massie, Matthew L., Brent N. Chun, and David E. Culler. "The ganglia distributed monitoring system: design, implementation, and experience." Parallel Computing 30.7 (2004): 817-840.

Design challenges

- Scalability
- Robustness
- Extensibility
- Manageability
- Portability
- Overhead
- Security*

Source: Massie, Matthew L., Brent N. Chun, and David E. Culler. "The ganglia distributed monitoring system: design, implementation, and experience." Parallel Computing 30.7 (2004): 817-840.

* Zanikolas, Serafeim, and Rizos Sakellariou. "A taxonomy of grid monitoring systems." *Future Generation Computer Systems* 21.1 (2005): 163-188.

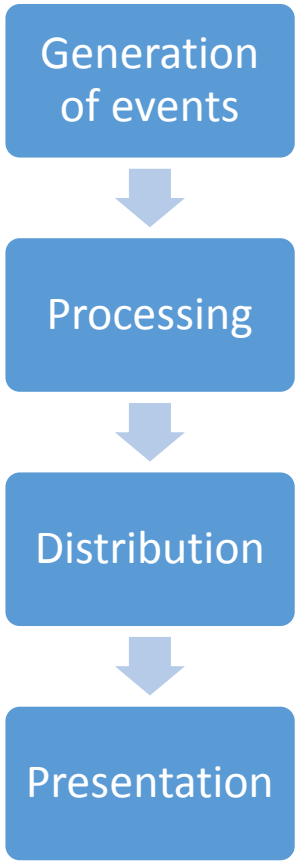
Sources of Event Data

- Sensor (for h/w, s/w ,e.g. CPU, memory, SNMP)
- Application (Monitoring apps, e.g. NetLogger, Autopilot)
- Database (Archive)
- External system (e.g. weather service)

Tierney, Brian, et al. "A grid monitoring architecture." (2002).

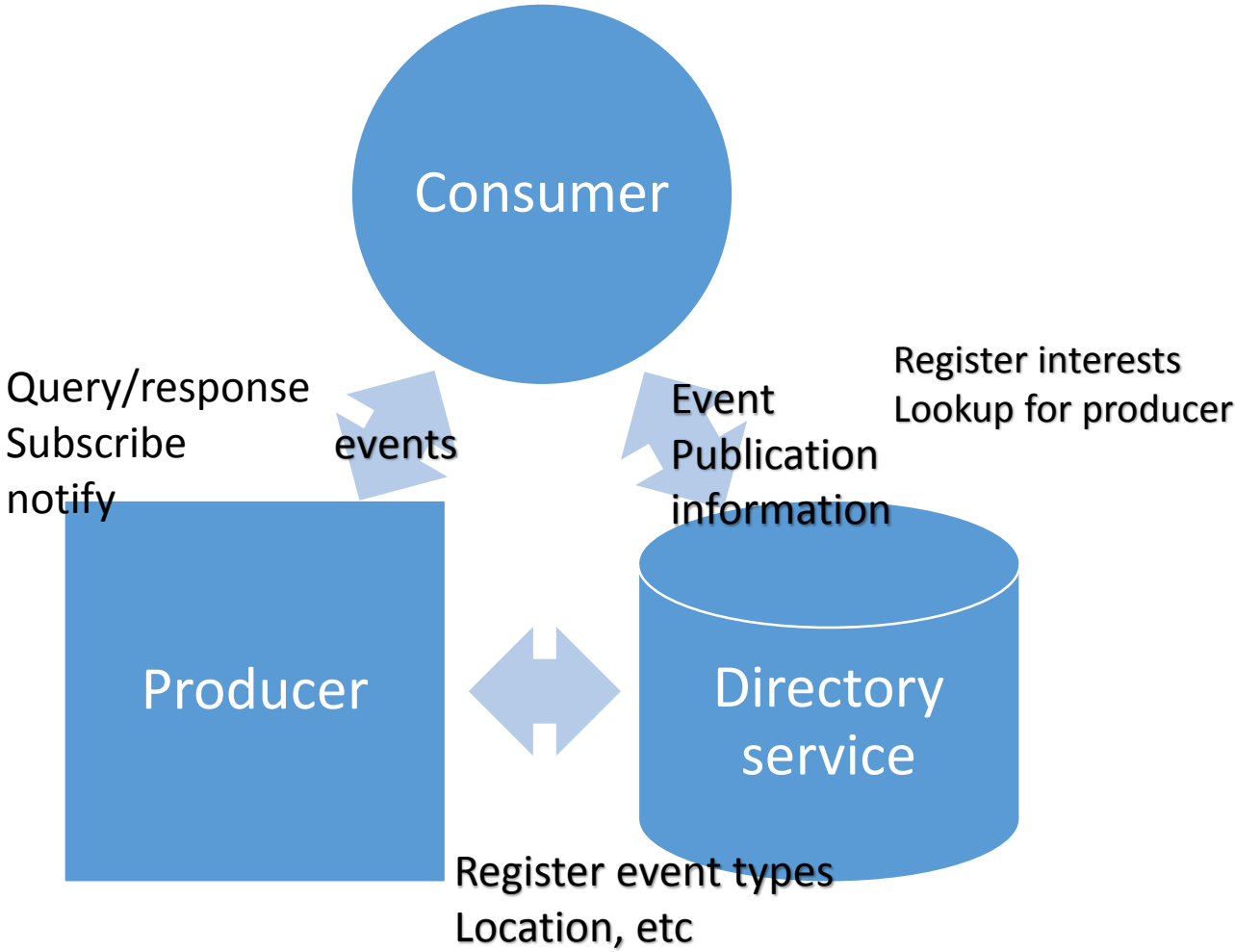
Monitoring process

- Sensors for the measurements
- Aggregating the data
- Delivery from source to destination
- Consumption (including visualization)



Grid Monitoring Architecture (GMA)

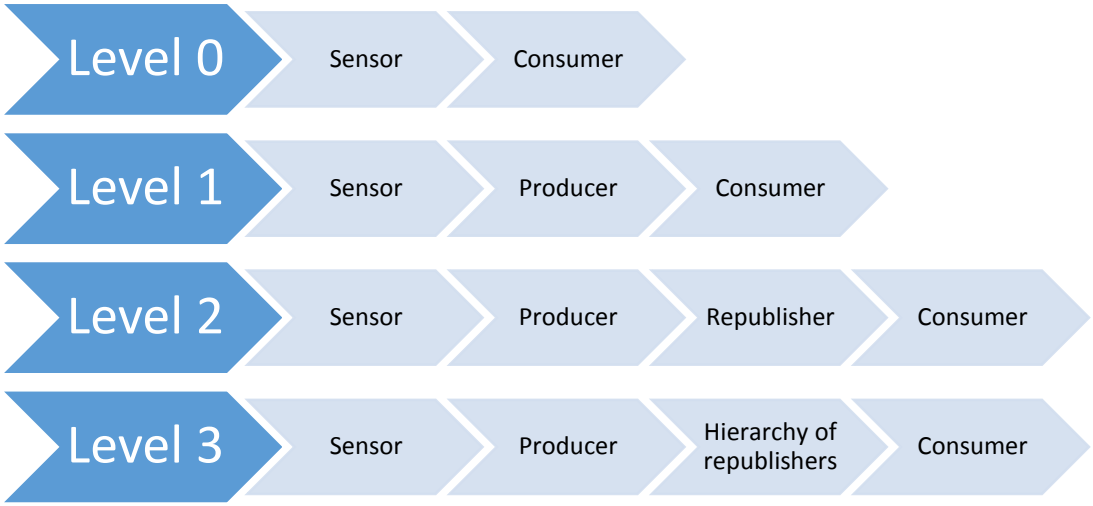
- Producer
  - Provides events
- Consumer
  - Receives events
- Directory service (Registry)
  - Lookup service (discovery)
  - Establish communication between consumer and producer



Tierney, Brian, et al. "A grid monitoring architecture." (2002).

Level of monitoring systems

- From GMA
  - Sensor
  - Producer
  - Consumer
- New components
  - Republisher
    - Processing, distribution
  - Hierarchy of republishers
    - More than one republisher



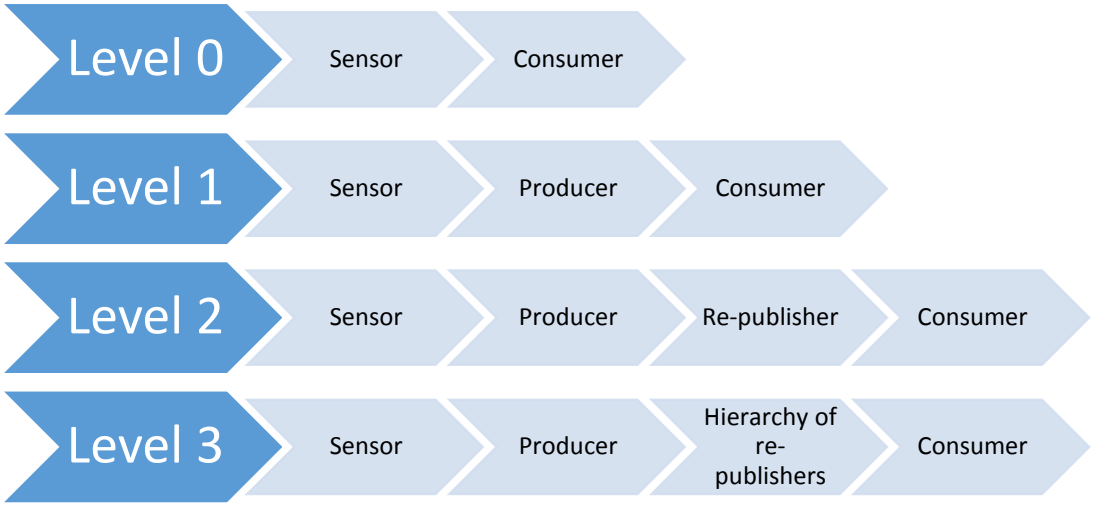
Taxonomy of monitoring systems

Zanikolas, Serafeim, and Rizos Sakellariou. "A taxonomy of grid monitoring systems." *Future Generation Computer Systems* 21.1 (2005): 163-188.

Level of monitoring systems

- Sensor
  - Generation of events
  - (Processing)
- Producer
  - (Generation of events)
  - (Processing)
  - Distribution
- Re-publisher
  - Processing
  - Distribution
- Hierarchy of Re-publishers
  - More than on processing and distribution
- Consumer
  - (Processing)
  - Presentation/Consumption

() is optional

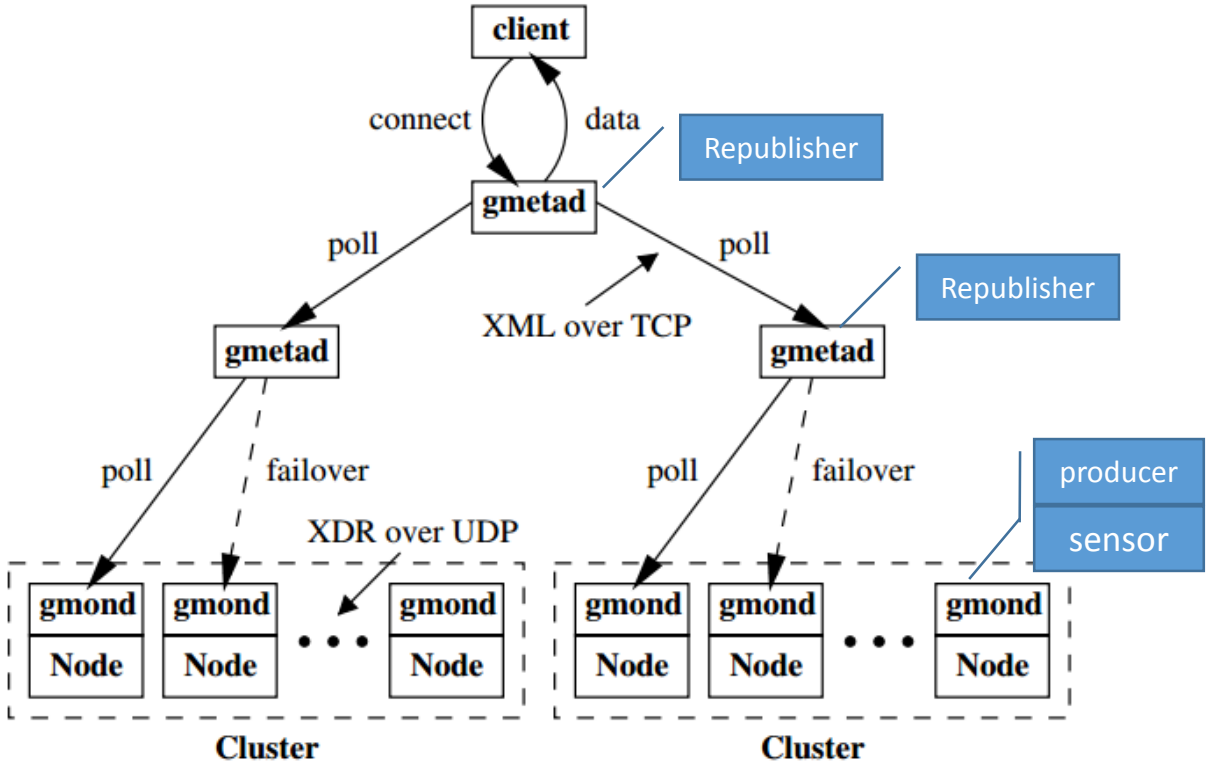


Taxonomy of monitoring systems

Zanikolas, Serafeim, and Rizos Sakellariou. "A taxonomy of grid monitoring systems." *Future Generation Computer Systems* 21.1 (2005): 163-188.

Example. Ganglia Distributed Monitoring System

- Level 3
  - Sensor, producer (gmond)
  - Hierarchy of Republishers (tree of gmetad)
  - Consumer (client)

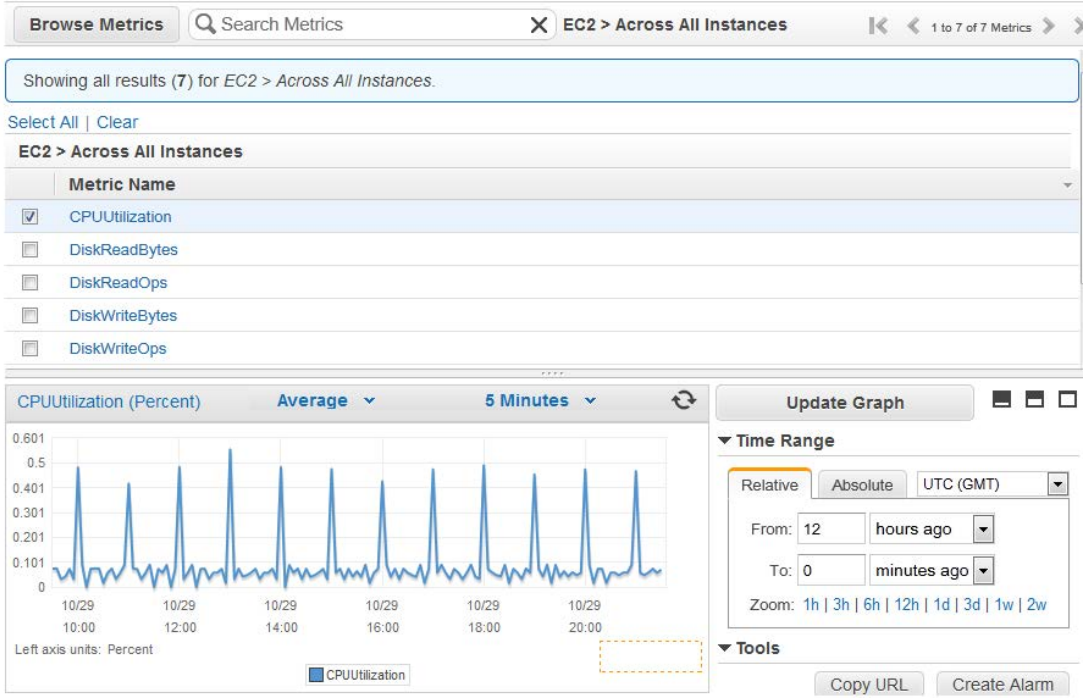


Ganglia Architecture

Massie, Matthew L., Brent N. Chun, and David E. Culler. "The ganglia distributed monitoring system: design, implementation, and experience."*Parallel Computing* 30.7 (2004): 817-840.

Monitoring in the cloud

- Complexity of the infrastructure increased
  - Shared resources in virtualization
  - Different service models (IaaS, SaaS, PaaS)
  - Data center
  - Public service
  - Billing / Accounting / Auditing / Profiling
- Grid Monitoring System modified for the cloud (bare-metal)
  - Plugin/add-on to grid monitoring applications (e.g. Eucalyptus with Nagios)
- Monitoring data from the hypervisor

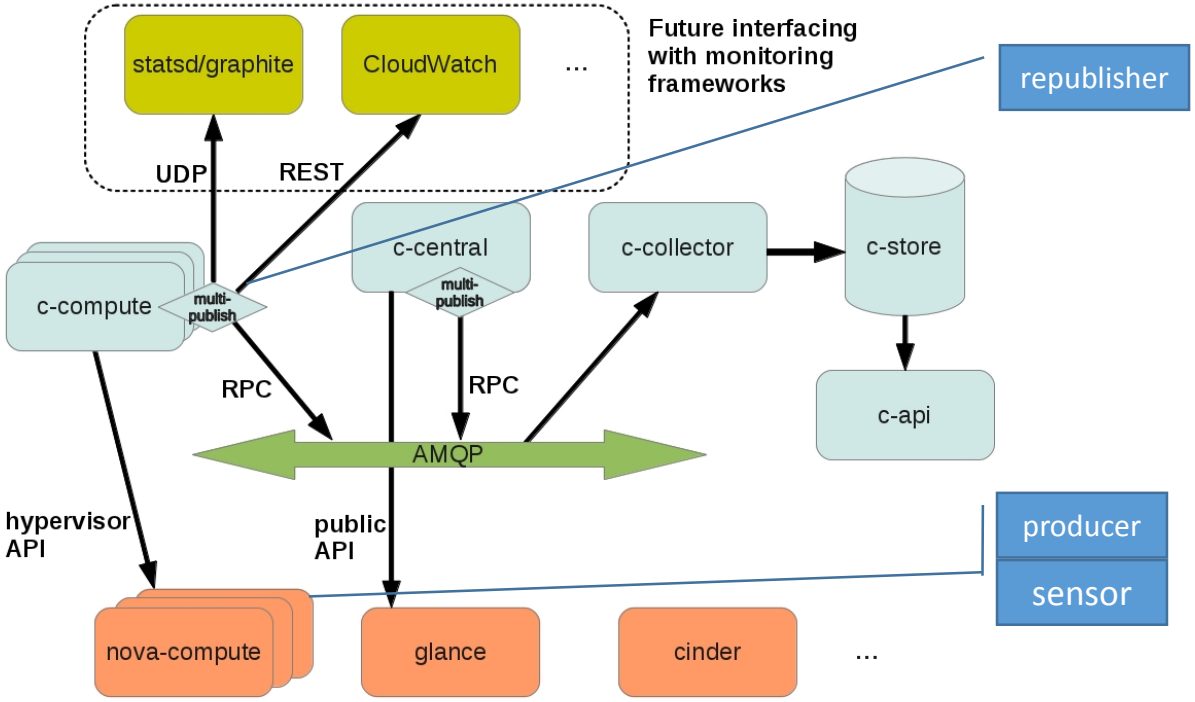


Amazon CloudWatch screenshot

Image source: aws.amazon.com

OpenStack Ceilometer

- Hypervisor (nova-compute) provides performance data and resource allocation data
- Billing system*
  - IaaS
    - Number of VMs
    - Size of CPUs, Memories, Disks (flavors)
  - PaaS
    - Task completion time
  - SaaS
    - Application-specific performance levels, functions



OpenStack Cloud with Monitoring

Image source: openstack.org

*Aceto, Giuseppe, et al. "Cloud monitoring: A survey." *Computer Networks* 57.9 (2013): 2093-2115.

Summary of Monitoring Distributed System

- Architecture
  - GMA by Global Grid Forum
- Taxonomy
  - Level 0 - 4
- Extension for the cloud
  - Work with hypervisor

Bioinformatics Applications in The Cloud

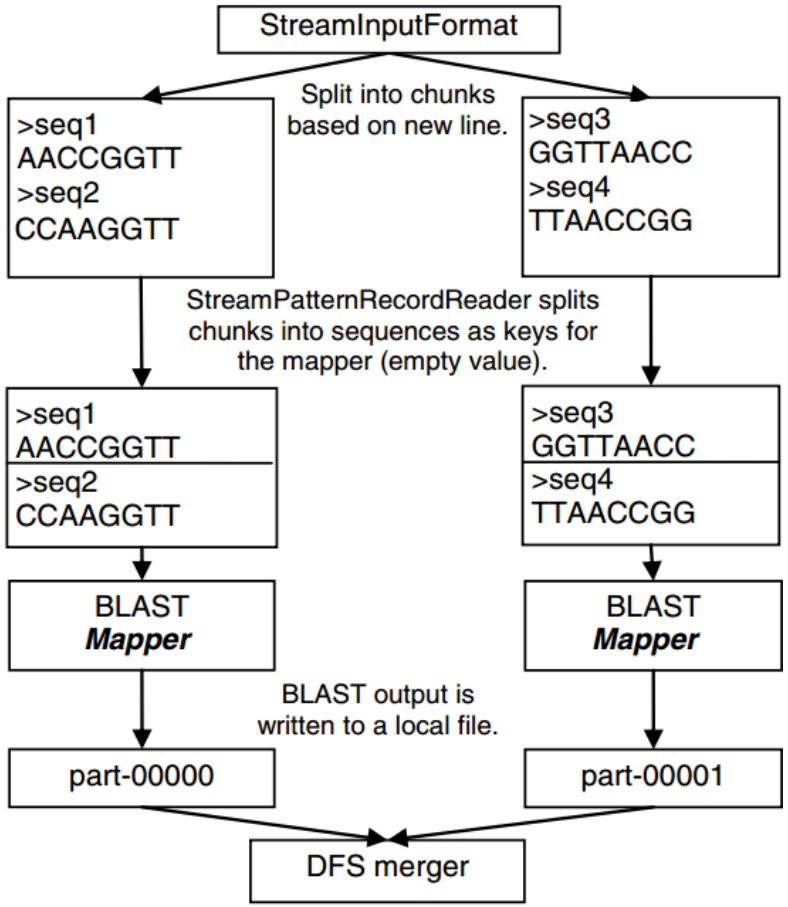
Parallel Processing and Environments

Outline

- Related work
  - Parallel processing (Hadoop/MapReduce)
  - Scientific workflow

Parallel Processing

- MapReduce
  - Independent separated tasks
- Cloud cluster
  - resizable

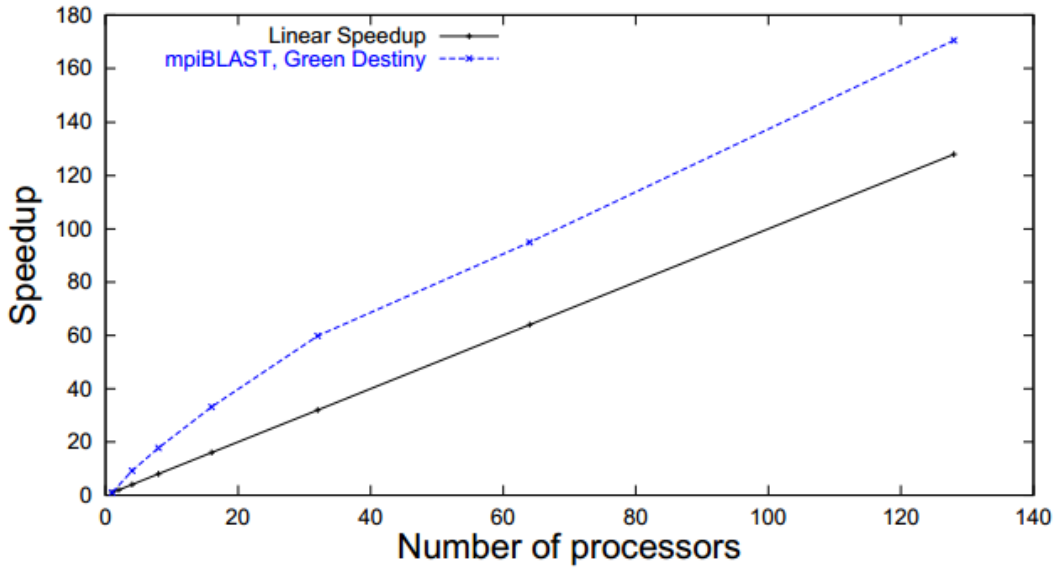


Example of MapReduce in BLAST search

Image source: Matsunaga, Andréa, Maurício Tsugawa, and José Fortes. "Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications." *eScience, 2008. eScience'08. IEEE Fourth International Conference on*. IEEE, 2008.

mpiBLAST

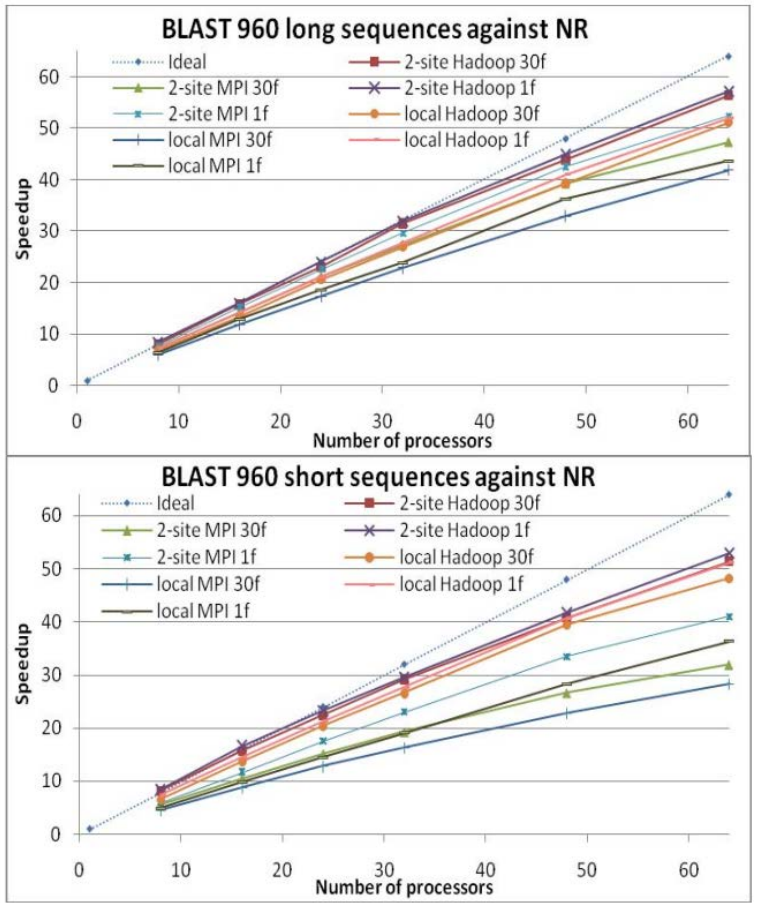
- BLAST runs a search algorithm with a database segmentation on Beowulf cluster
- Database segmentation using Messaging Passing Interface (MPI)
- No fault tolerance



Speedup of mpiBLAST
(300kb query sequences, 5.1GB database)

CloudBLAST

- Better performance than mpiBLAST (not significant)
- Failure recovery by MapReduce
- Xen VMs on two regions

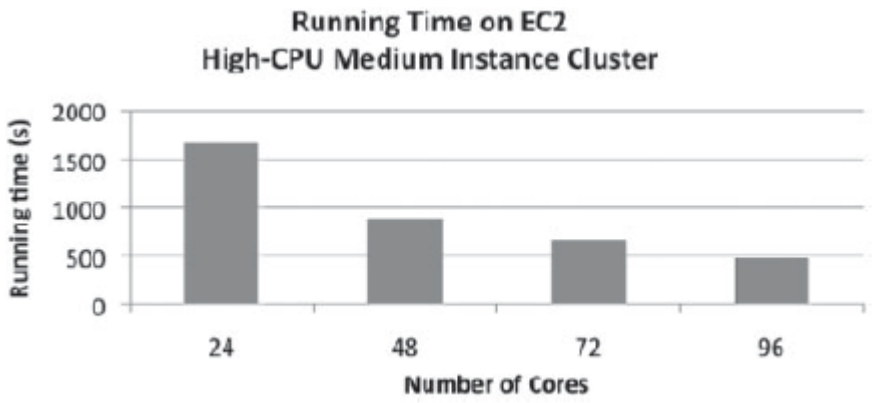


Speedup curves for CloudBLAST (Hadoop) and mpiBLAST

Image source: Matsunaga, Andréa, Maurício Tsugawa, and José Fortes. "Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications." *eScience, 2008. eScience'08. IEEE Fourth International Conference on*. IEEE, 2008

CloudBurst

- Is a read mapping algorithm using map() and reduce() functions
- Runs on Amazon EC2
- 30x faster than RMAP (short-read mapping software)
- Better performance with more vCPUs (32 bit dual core 3.2 GHz Intel Xeon)
- 7M short reads to human genome (3Gbp)

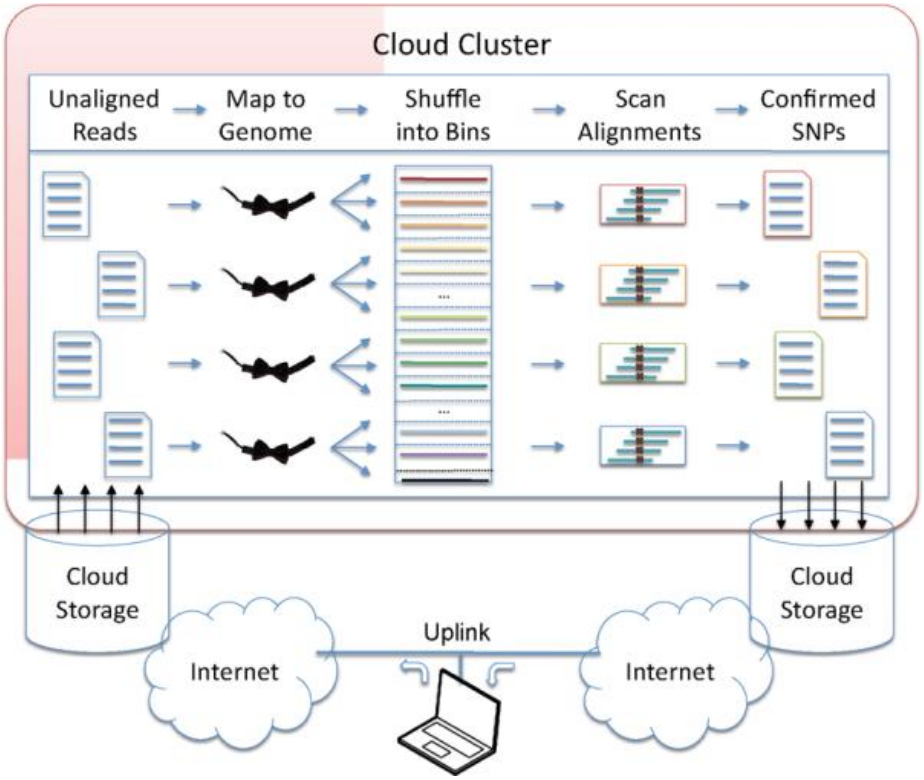


Comparison of CloudBurst running time on EC2

Image source: Schatz, Michael C. "CloudBurst: highly sensitive read mapping with MapReduce." *Bioinformatics* 25.11 (2009): 1363-1369.

Crossbow

- Genotyping program using Bowtie, SOAPsnp and Hadoop
- Larger input data with more compute resources than CloudBurst
- 2.7 billion reads – 103 GB
  - (385x bigger than CloudBurst)
- 320 vCPUs on 40 workers
  - Per 8 cores Xeon E5-2680 2.80 GHz
- Two issues
  - Input data transfer to the cloud
  - Expertise to applying apps on Hadoop

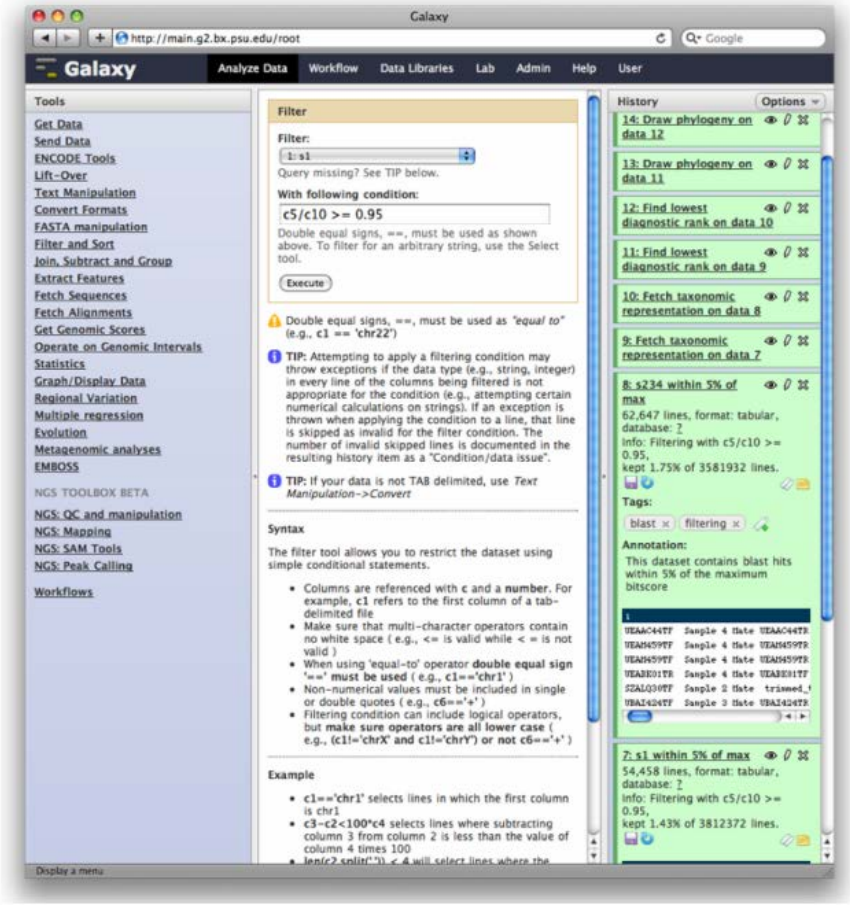


Map-Shuffle-Scan framework used by Crossbow

Image source: Schatz, Michael C., Ben Langmead, and Steven L. Salzberg. "Cloud computing and the DNA data race." *Nature biotechnology* 28.7 (2010): 691.

Scientific workflow

- Usability
- Supports parallel programming framework
- Provides data analysis environments
- Works with cloud platforms

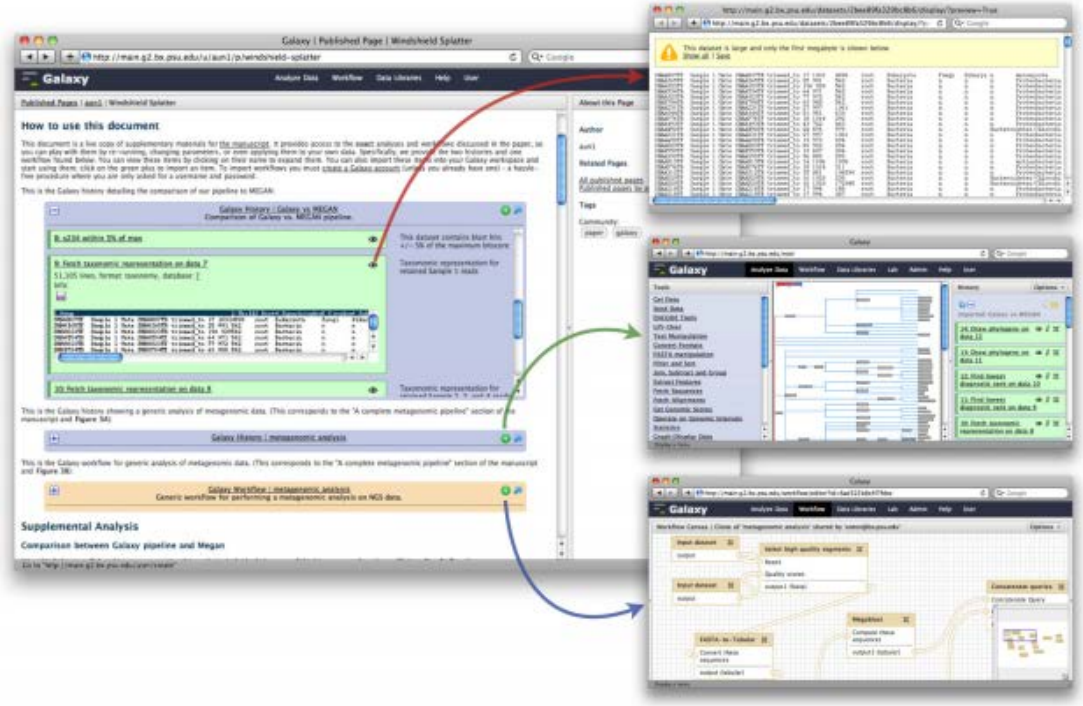


Galaxy workflow

Image source: Goecks, Jeremy, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biol* 11.8 (2010): R86.

Galaxy Workflow System

- Accessibility
  - Public web service
  - Data import (local or data warehouse e.g. UCSC)
  - ToolShed (software repository)
- Reproducibility
  - Recorded workload
- Transparency
  - Public repository to share experiments and tools
- CloudMan runs Galaxy on Amazon EC2
- Crossbow, CloudBurst on Galaxy tools

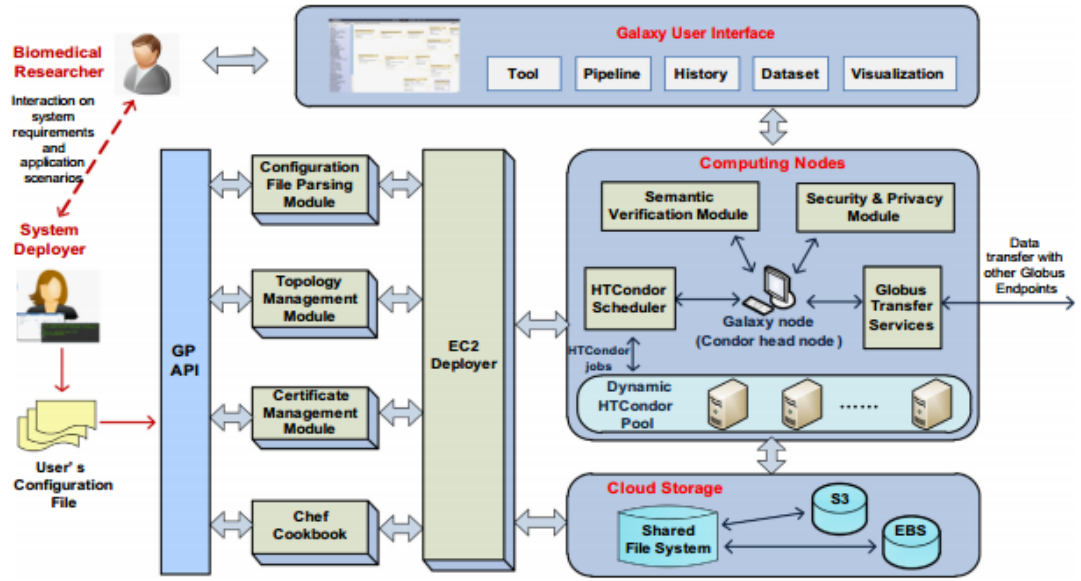


Galaxy pages

Image source: Goecks, Jeremy, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biol* 11.8 (2010): R86.

Galaxy on the Cloud

- Cloud Access
  - CloudMan
    - Galaxy Compute Cluster
  - CloudBioLinux
    - Suite of bioinfomatics software
  - Amazon EC2, Eucalyptus, or VirtualBox
- Data Transfer
  - Globus Transfer (GridFTP)
  - Cloud Storage (like Amazon Public Data Sets)



Architecture of Cloud-based bioinformatics workflow platform

Image source: Liu, Bo, et al. "Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses." *Journal of biomedical informatics*(2014).

Cloud-enabled bioinformatics platforms

Name	Year	Description	Application tools
CloudBLAST	2008	Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications	Hadoop, ViNe, BLAST
CloudBurst	2009	highly sensitive read mapping with MapReduce	MapReduce, Amazon EC2
Crossbow	2009	Searching for SNPs with cloud computing	Hadoop, bowtie, SOAPsnp, Amazon EC2
Myrna	2010	Cloud-scale RNA-sequencing differentialexpression analysis	Hadoop, Amazon EMR, HapMap
Galaxy	2010	Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences	Python, web server, SQL database
Galaxy CloudMan	2010	delivering cloud compute clusters	Amazon EC2, Bio-Linux, Galaxy
AzureBlast	2010	A Case Study of Developing Science Applications on the Cloud	Azure, BLAST
CloudAligner	2011	A fast and full-featured MapReduce based tool for sequence mapping	CloudBurst, MapReduce, Amazon EMR
CloVR	2011	virtual machine for automated and portable sequence analysis from the desktop using cloud computing	VM, VirtualBox, VMWare
Cloud BioLinux	2012	pre-configured and on-demand bioinformatics computing for the genomics community	VM, Amazon EC2, Eucalyptus, VirtualBox
FX	2012	an RNA-Seq analysis tool on the cloud	Hadoop, Amazon EC2
Rainbow	2013	Tool for large-scale whole-genome sequencing data analysis using cloud computing	Crossbow, bowtie, SOAPsnp, Picard, Perl, MapReduce
BioPig	2013	a Hadoop-based analytic toolkit for large-scale sequence data	Hadoop, Apache Pig
SeqPig	2014	simple and scalable scripting for large sequencing data sets in Hadoop	Hadoop, Apache Pig
SparkSeq	2014	fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision	Apache Spark, Scala, samtools

Summary of Bioinformatics Apps in The Cloud

- Parallel Processing
- Scientific Workflow

Topics

- Virtualization
- Monitoring Distributed Systems
- Bioinformatics Applications in The Cloud

Questions?

Thank You