How to Deploy OpenStack on Tianhe-2 Supercomputer

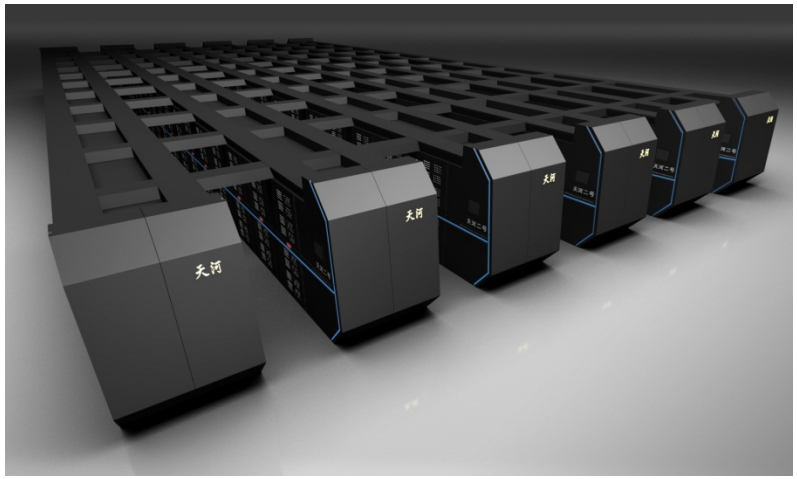
Yusong Tan
Associated Professor
National University of Defense Technology
November 5, 2013

Contents

- Background
- Deployment
- Optimization
- Preliminary Evaluation
- Contributions to Community
- Cases
- Next Steps

Background

- TH-2 Supercomputer
  - Sponsored by 863 High Tech. Program of Ministry of Science and Tech of China, Government of Guangdong province and Government of Guangzhou city
  - Built by National University of Defense Technology



中华人民共和国科学技术部 The Ministry of Science and Technology of the People's Republic of China

國防科學技術大學 NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

广东省人民政府 People's Government of Guangdong Province

中国广州政府 www.gz.gov.cn

Background

No. 1 of 41st Top 500 list at June 2013

Rpeak 54.9PFlops， Rmax 33.9PFlops

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National University of Defense Technology China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510
7	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.9	5,872.0	2,301
8	DOE/NNSA/LLNL United States	**Vulcan** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393,216	4,293.3	5,033.2	1,972
9	Leibniz Rechenzentrum Germany	**SuperMUC** - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM	147,456	2,897.0	3,185.1	3,423
10	National Supercomputing Center in Tianjin China	**Tianhe-1A** - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 NUDT	186,368	2,566.0	4,701.0	4,040

Background

- TH-2 Supercomputer
  - Neo-heterogeneous architecture
    - Xeon CPU & Xeon Phi @ X86 ISA

Items	Configuration
Nodes	16000
Processors	32000 Intel Xeon CPUs + 48000 Xeon Phis + 4096 FT CPUs
Interconnect	Proprietary high-speed interconnection network TH-NI
Memory	1PB in total
Storage	Global shared parallel storage system, 12.4PB
Cabinets	162=125+13+24 compute/communication/storage Cabinets
Cooling	Closed Air cooling system
Performance	Peak performance is 54.9PFlops Max performance is 33.9PFlops

Background

- TH-2 Supercomputer
  - Installed in National Supercomputer Center in Guangzhou (NSCC-GZ)
    - Open platform for research and education
      - To provide HPC service
    - Public information infrastructure
      - To accelerate the industry and economy

More than HPC!

Background

- TH-2 Supercomputer
  - How to be used as a public information infrastructure

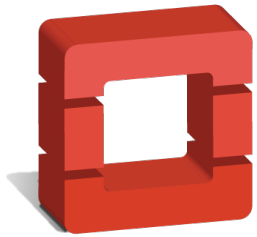
    Hybrid Cloud infrastructure
    - Virtualization based IaaS cloud computing!
    - Different environments
      - Redhat, CentOS, Windows, Ubuntu, SUSE…
    - Different scales
      - From several nodes to thousands of nodes

Background

- Hardware for OpenStack deployment
  - 6400 nodes on TH-2
    - 50 cabinets
    - 128 nodes in each cabinet
  - Each node:
    - Intel Xeon Ivy Bridge (12 cores) * 2
    - 96 GB RAM
    - 1TB local disk
    - GE Nic * 2
    - TH-NI High-speed Network (160 Gbps)
    - Intel Xeon Phi * 3

Background

- Software



openstack™
CLOUD SOFTWARE

Grizzly

ubuntu®

Server 12.04.2 LTS

ceph

v0.56.6

puppet labs

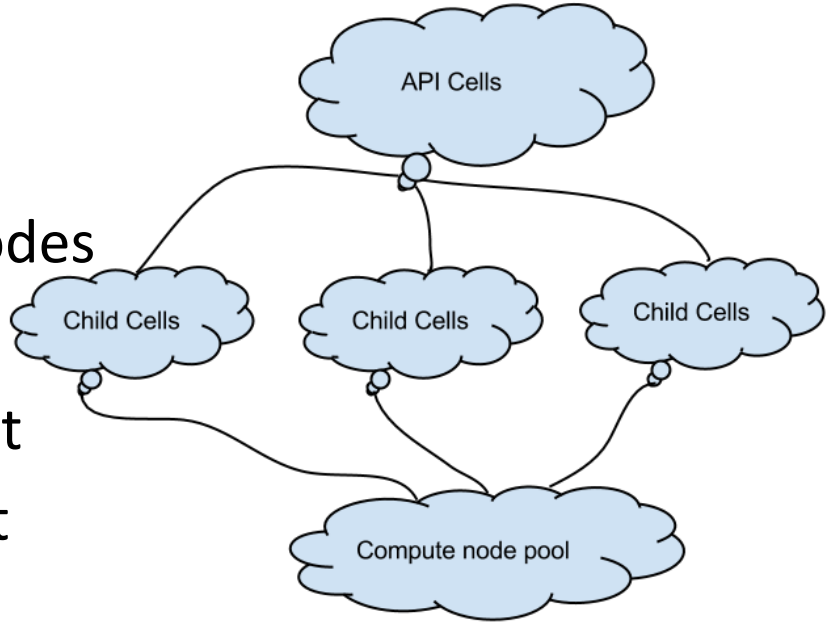
v2.7.11

Contents

- Background
- Deployment
- Optimization
- Preliminary Evaluation
- Contributions to Community
- Cases
- Next Steps

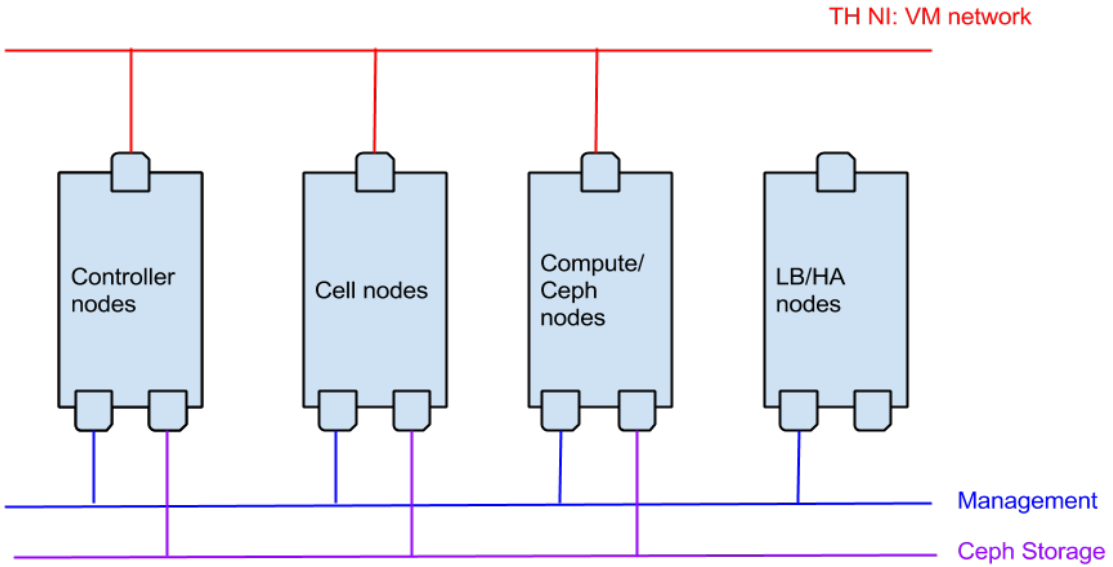
Deployment

- Architecture
  - 256 controller nodes
    - 2 cabinets
    - 124 api nodes: nova-api, nova-cells, glance-*, cinder-api, cinder-scheduler, neutron-server, keystone
    - 4 LVS+Keepalived
    - 96 network nodes
    - 32 Ganglia+Nagios server nodes
  - 3840 compute nodes
    - 30 cells, each for one cabinet
    - 2 cell servers in each cabinet



Deployment

- Network Topology
  - 1 GE for management
  - 1 GE for Ceph storage
    - higher performance with TH-NI virtualized Ethernet
  - TH-NI virtualized Ethernet for VMs communication



Deployment

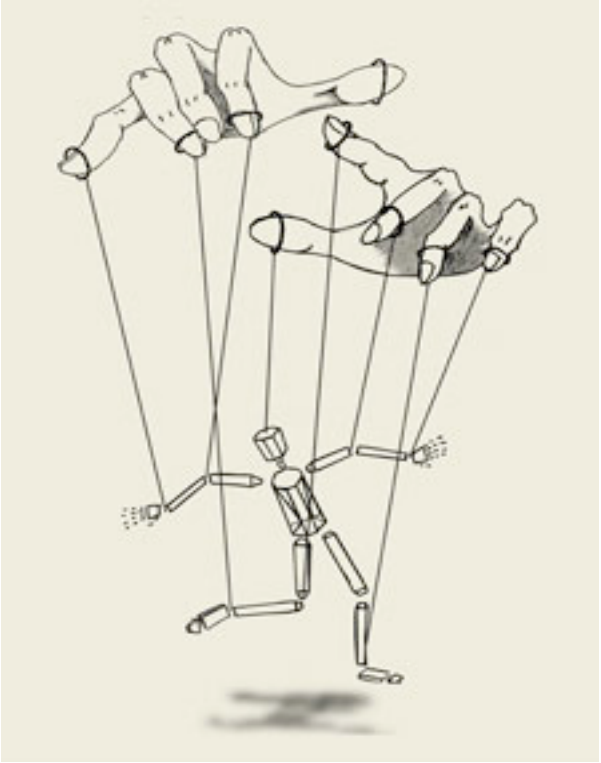
- Storage

  Ceph as “all-in-one”

  - RBD for Glance image storage
  - RBD for volume backend
  - FS for instance storage

Deployment

- Puppet
  - Different manifests for different roles
  - Each node reads its role from same conf file



Deployment

- Implementation
  - Ubuntu Server 12.04.2 LTS based diskless system
  - Configure the role of each node using Puppet when startup
  - Using TH-NI to pull images to accelerate booting
    - With native TH-NI RDMA driver support
  - Other configurations
    - Partition hard disks when first boot
    - SSH without password
    - Decide IP and hostname based on TH-NI's nid dynamically

Contents

- Background
- Deployment
- Optimization
- Preliminary Evaluation
- Contributions to Community
- Cases
- Next Steps

Optimization

- More nodes
  - Load balance for *-apis
  - Separated Keystone API service for Neutron

Optimization

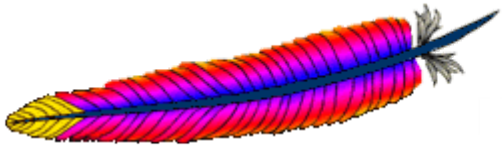
- More workers

Optimization

- More workers
  - Multi-processed
    - Multi-processed neutron-server
      - api_workers
    - Multi-processed nova-api
      - ec2_workers
      - osapi_compute_workers
      - metadata_workers
    - Multi-processed nova-conductor
      - workers
    - Multi-processed glance-api
      - workers

Optimization

- More workers
  - Apache hosting
    - Keystone
    - supported by its design and implementation



Apache

Optimization

- More powerful
  - Nova
    - Eliminate api rate limit
      - api_rate_limit=false
    - Use large db pool size
      - sql_dbpool_enable
      - sql_min_pool_size
      - sql_max_pool_size
      - sql_max_overflow

Optimization

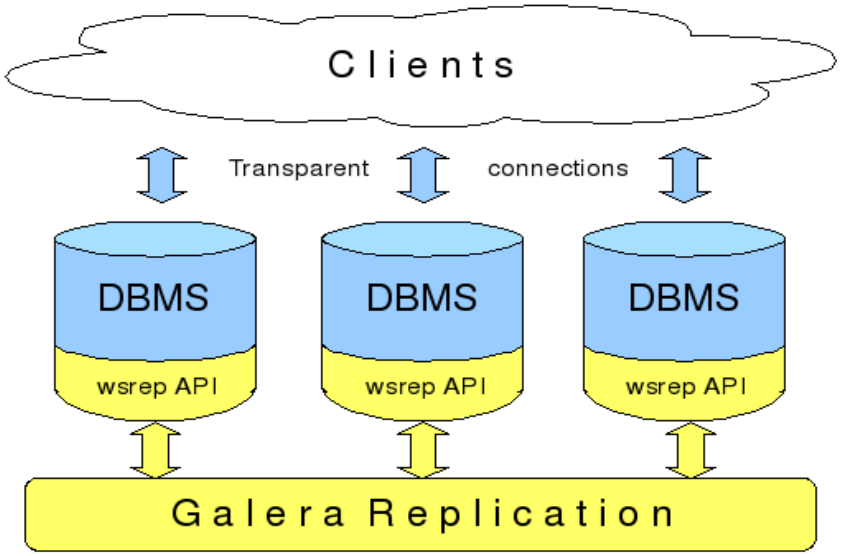
- More powerful
  - Neutron
    - Use larger db pool size
      - sqlalchemy_pool_size
    - Increase agent down time
      - agent_down_time

Optimization

- More powerful
  - Rabbitmq
    - Set high memory watermark
      - rabbitmqctl set_vm_memory_high_watermark 0.6
    - Set high socket limit
      - ulimit -n 102400
    - Rabbitmq Cluster
      - Clustering rabbitmq by running more rabbitmq-servers to distribute workload

Optimization

- More powerful
  - MySQL
    - Use larger maximum connections
      - max_connections=102400
    - Galera Mysql cluster
      - Chose some api servers run one mysql master
      - using LVS to provide LB
      - access db via a single entry point by using a virtual IP



Optimization

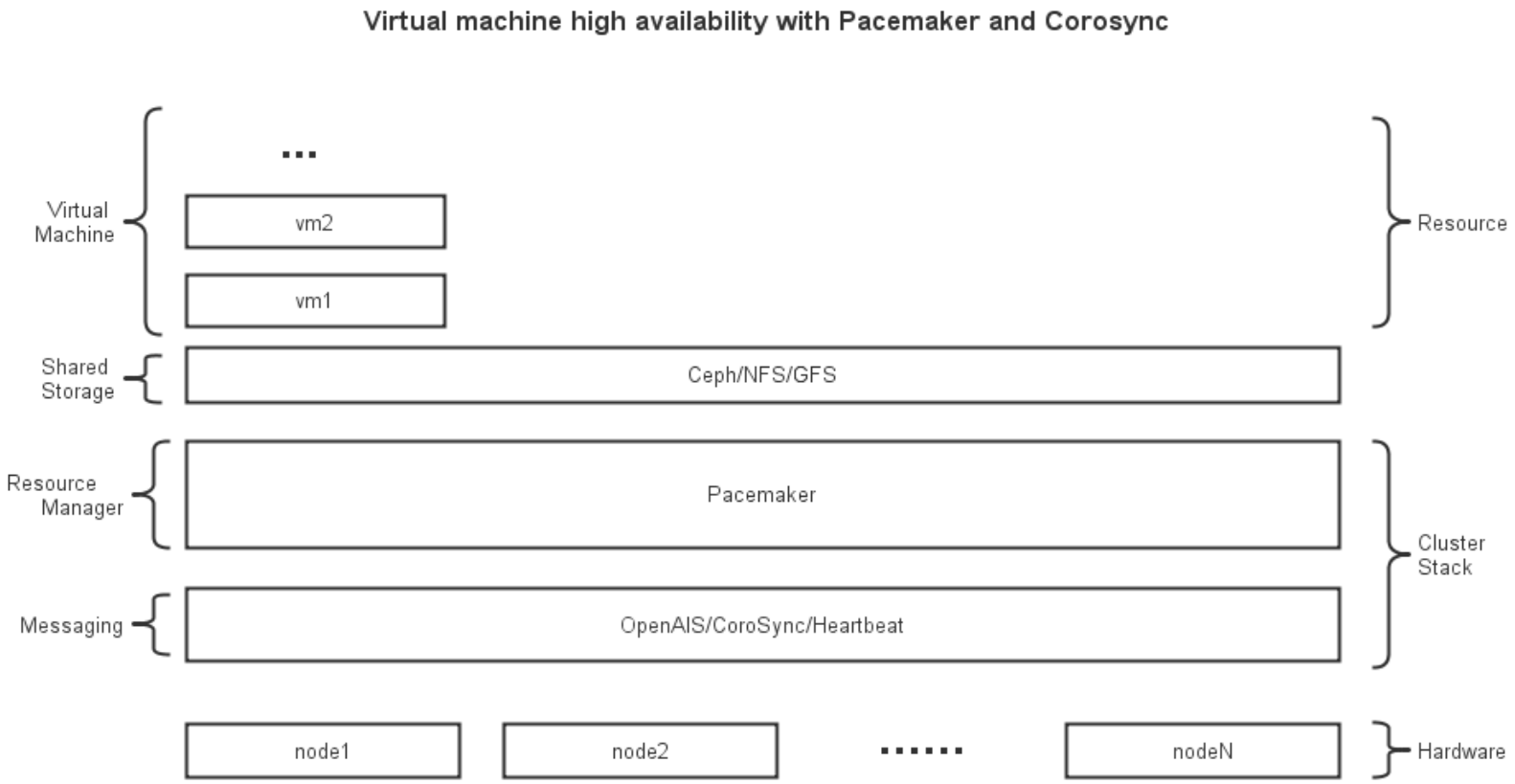
- More powerful
  - KVM
    - Huge pages enabled
    - Using VhostNet
      - provides better latency and greater throughput
    - Using virtio_blk
      - Get higher performance of storage devices
    - Kernel SamePage merging enabled(KSM)
      - Because we want to oversell
    - Kernel I/O scheduler to "Deadline"

Optimization

- Services HA/LB
  - LVS + keepalived for all API servers
  - Increased rpc_response_timeout
  - Increased quantum_url_timeout

Optimization

- HA for instances



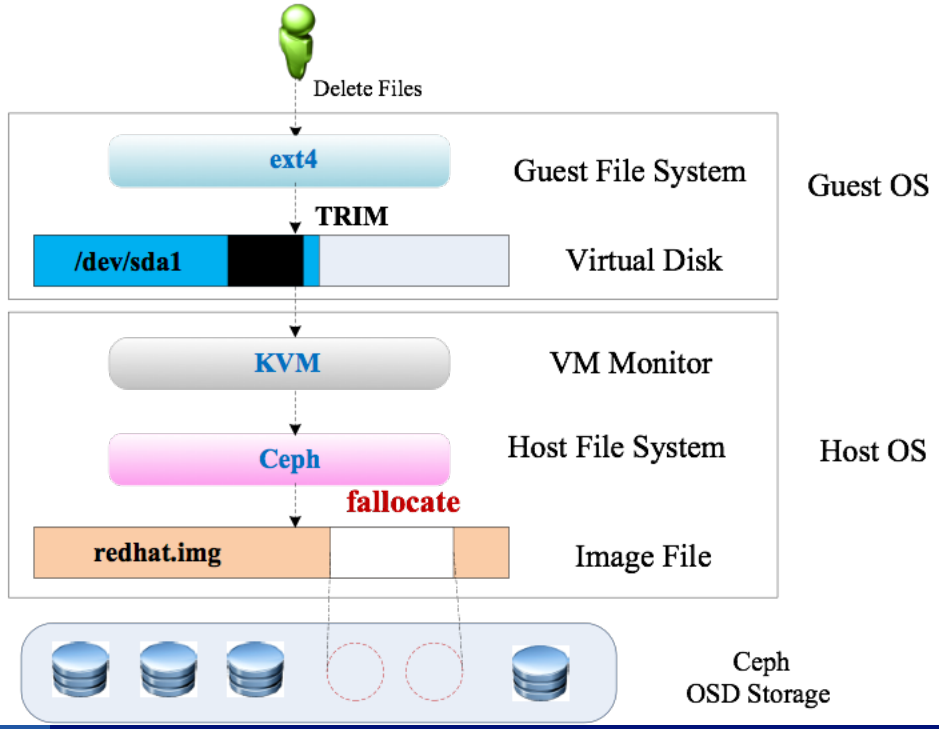
Optimization

- Ceph
  - Inline data support for small file IO
  - Typical file read/write traffic
    - Client consult MDS for file metadata
    - Client communicate with OSD for file data

  - With inline data support, for small files
    - Client consult MDS for both

Optimization

- Ceph
  - Punch hole support
    - Enable sparse file support
    - Improves space efficiency in virtualization situation



Contents

- Background
- Deployment
- Optimization
- Preliminary Evaluation
- Contributions to Community
- Cases
- Next Steps

Preliminary Evaluation

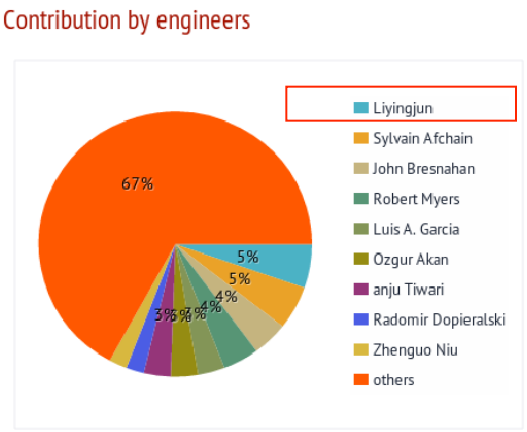
- Maximum VMs started one round
  - cirros test image, m1.tiny
  - ~5300
  - scheduler cost most time
- Concurrent request handled by api server per second
  - ~1100 list instances operations (active instances num. =1000)
  - Bottlenecks
    - GE, use TH-NI instead
    - nova-api (if TH-NI is used), use more api servers
      - Time spent on query of database only takes less 40%

Contents

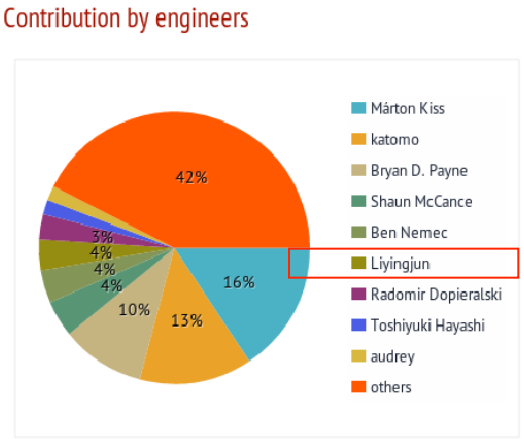
- Background
- Deployment
- Optimization
- Preliminary Evaluation
- Contributions to Community
- Cases
- Next Steps

Contributions to Community

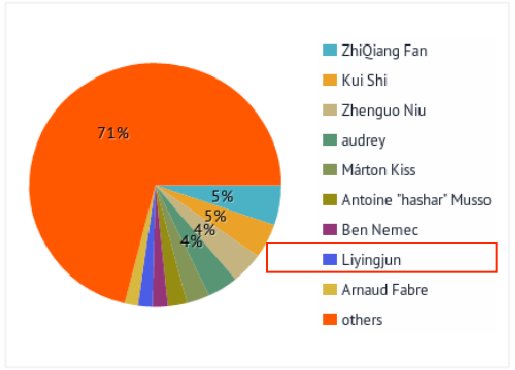
- For Havanna
  - 5 blueprints, 10 bug-fixs is merged (3976 LoC, till 2013.10.12)
  - ranked 1 (BP)/6 (LoC) /9 (Commit) of independent developers



BP



LoC



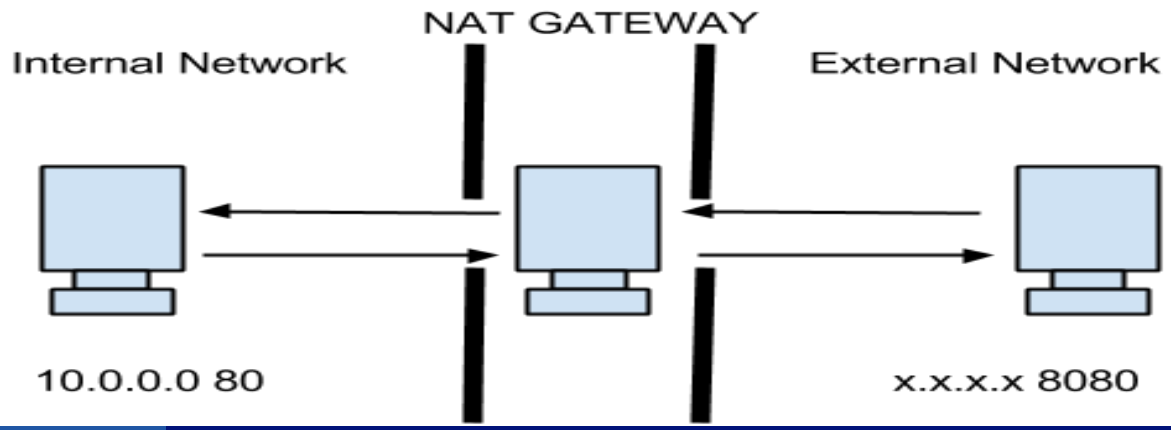
Commit

Contributions to Community

- Quota
  - per-project-user-quotas-support
    - The administrators can set quotas on a per-project-user basis, rather than just a per-project basis.
  - editable-default-quotas-support
    - The administrators can edit the default quota for nova or cinder through horizon or CLI.

Contributions to Community

- Access VMs via port mapping
  - Existing approaches to access VMs
    - Floating IP: limited by the number of available IPs
    - VPN: relatively complex to configure
  - Our solution
    - Unused ports of the public IP can be dynamically mapped to different VMs to supply a access path



Contributions to Community

- New UI design & feature enhancements for Horizon

Contributions to Community

- For Ceph
  - Two blueprints (Inline data support for small file IO+ Punch hole support) presented at the first Ceph Developer Summit
  - Both are pushed into upstream
  - 20 commits, 1000+ LOC, including 600+ for linux kernel

Contributions to Community

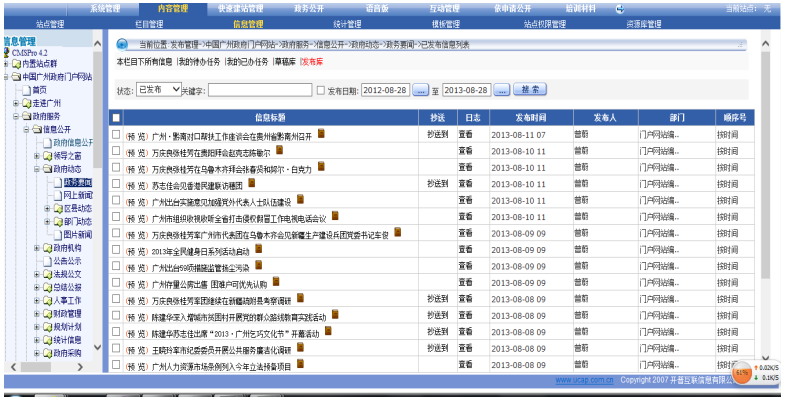
- And More
  - Add Multi-level User Management to Keystone via Project Nesting (in development)
  - Multiple Level User Quota Management (in development)
  - More bugs fixed
  - …

Contents

- Background
- Deployment
- Optimization
- Preliminary Evaluation
- Contributions to Community
- Cases
- Next Steps

Cases

- E-Gov for Guangzhou City
  - Websites of Guangzhou government
    - Requirements： Web/Application/Database Servers
    - Efforts:
      - Load Balance
      - High Availability



Cases

- E-Gov for Guangzhou City
  - E-Gov Data Management System
    - Requirements： JBoss/WebLogic/Oracle Servers
    - Efforts
      - Oracle @ Ceph RBD
      - VM Optimize for Oracle & Java

Contents

- Background
- Deployment
- Optimization
- Preliminary Evaluation
- Contributions to Community
- Cases
- Next Steps

Next Steps

For Supercomputer Center

- More performance tests
  - Performance of cell
  - Neutron
  - Glance
- Automatic Operation for large scale system
  - How to operate 6400 nodes
- Upgrade to Havana
- IaaS Services
  - More E-Government applications from Guangzhou
  - IaaS resource lease
- XaaS Service
  - Database, Hadoop, HA and more…..

Next Steps

For Community

- Automatic operation tools for large scale system
  - Deployment
  - Monitor
  - Metering & Billing
- Elastic resource management
  - QoS-based VM cluster scale-in/out
- Ceph
  - Read ahead optimization
  - SSD based writeback cache support

This is our Team!

Q&A