Argonne NATIONAL LABORATORY

Magellan: Building high-performance Openstack Clouds

Narayan Desai

OFA User Day 2013

Mathematics and Computer Science Division

Argonne National Laboratory

U.S. DEPARTMENT OF ENERGY

Talk Goals

- Describe high level cloud architecture
- Survey network use cases
  - Particularly how these differ from HPC
- Describe the good, the bad, and the ugly

What are clouds?

- Infrastructure as a Service
- Platform as a Service
- Software as a Service
- Big Data
- MP3s (?)
- Storage Services
- Music Streaming
- Photo Hosting
- Hosted Blogging Platforms
- Social Networking
- Hyperscale Data Centers
- Security Services
- Application Auto-scaling
- Helpdesk as a Service
- Etc

The Private Cloud Software Stack

- Amazon EC2 is the archetype
- Provides a basic IaaS model
  - Compute resources (VMs)
  - Several flavors of CPU/memory/ephemeral disk configurations available
  - VM Image repository
  - Persistent block storage
  - Object storage
  - Private networks with public outbound connectivity
  - Bridged layer 2 networks
  - Security policy
- All of these resources are provided behind APIs
  - Disintermediated
  - Programs can control resource allocation
- Multi-tenant (via security policy)
- Designed initially for interactive workloads, but can support batch

Magellan History

- Initially funded by DOE ASCR/ARRA
- Built a moderate sized cloud testbed
  - IBM Gear, 7500 cores, 30 TB of memory, 1.5 PB of storage
  - Mellanox QDR Infiniband
  - Connected to ESNet/ANI, now transitioning to ESNet 5
- First system to run Openstack at real scale
  - Largest deployed system worldwide from early 2011 until mid 2012
- Transitioned to support the DOE KBase project in late 2012
  - Mixed workload
    - Data analysis
    - Compute intensive
    - Web service APIs
- Research goals
  - Support wide range of workloads ill-suited for supercomputers
  - Narrow performance gap caused by virtualization

Openstack

- Openstack is an open-source cloud software stack
  - Born out of frustration with Eucalyptus, actually
  - NASA experienced problems on their Nebula system, and decided to write their own cloud software stack
  - Early partnership with Rackspace
  - Active community (2600 (!) people in PDX this week)
- Built around an SQL DB and message bus
- Openstack evaluated best when we compared options (based on the Bexar release)
- Deployment was fairly straightforward and successful
  - Stable and scalable system
  - Didn't need to worry about high fault rates

So how does Openstack use networks, anyway?

- System control plane
- Block Storage
- Inter Virtual Machine Networking
- VM to internet traffic

Openstack Control Plane

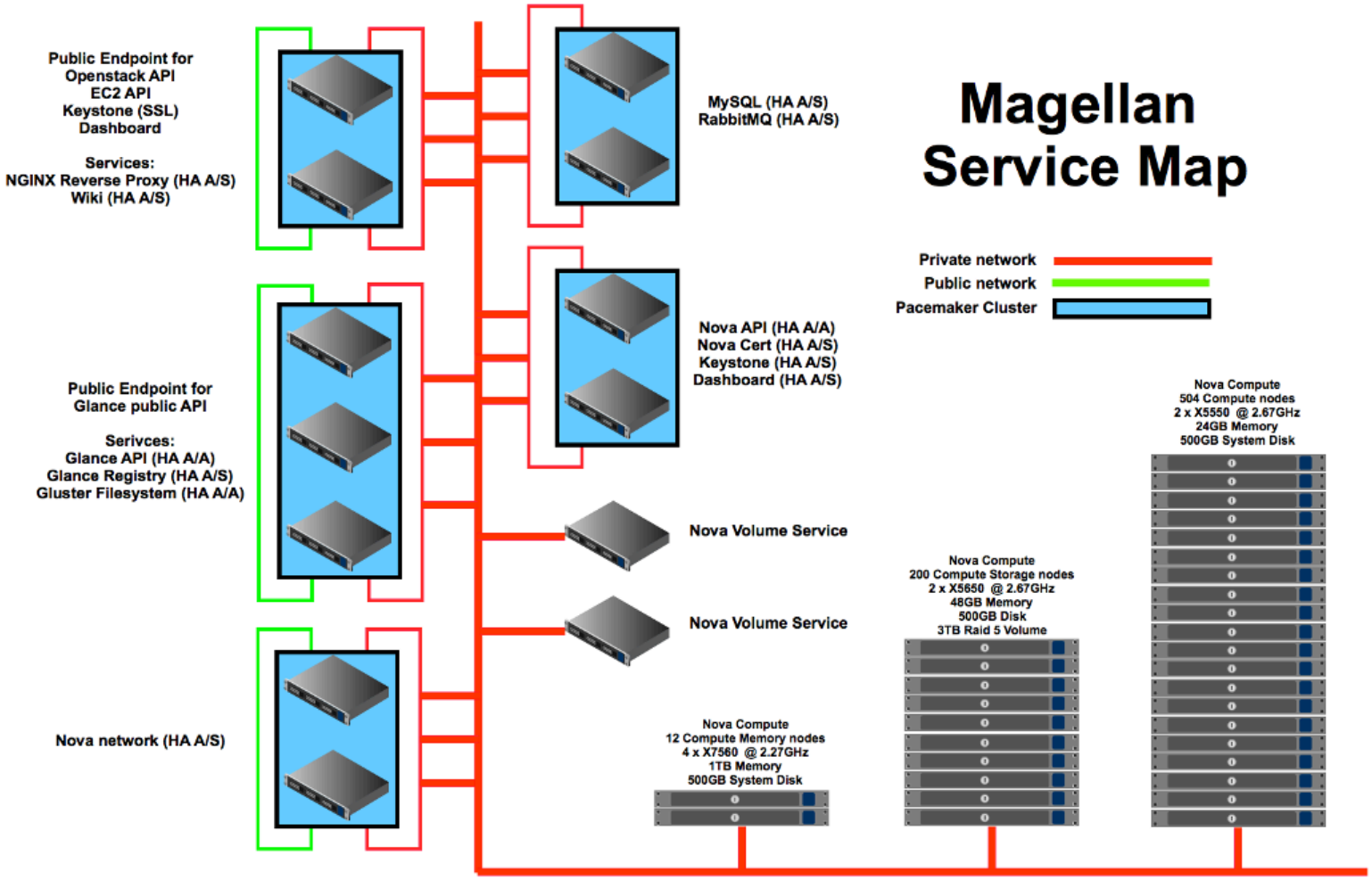
Provides all resource management and orchestration services

- Load Balancers
- Shared IP addresses
- MySQL
- RabbitMQ
- HTTP APIs
- Glance (OS Images)

Observations:

- All TCP/IP services
- Easily run over IPoIB, for the most part
- High Availability IP addresses are the only tricky part, and not even all that hard
- Running the control plane over IB provides considerably better scalability than others see with similar systems on 1GE/10GE ethernet

Magellan Software Architecture



Block Storage

Provides access to storage via a block device interface

- Terminated in the hypervisor
  - (No need for virtualized drivers)
- Variety of protocols can be used here
  - Parallel/Distributed filesystem (via VFS)
    - Any of the usual suspects
  - Distributed object storage (via Qemu driver)
    - GlusterFS/Ceph
  - SRP
  - ISCSI

Block Device Experiences

Today

- ISCSI for block devices
  - Was well supported by Openstack
  - Good target implementation in Comstar
  - Bad baseline performance over IPoIB
    - ~300 MB/s
    - Tuning nightmare
  - iSER considerably better
    - ~1800 MB/s
    - No tuning needed

Soon

- GlusterFS for converged storage
  - Block/Object/Filesystem
  - Multi-tenant
  - Native RDMA support
  - Qemu block driver
  - Replication-based fault tolerance
  - Enables migration
- Performing initial PoC now

Technical Workload Evaluation

- As we expected, tightly coupled workloads were a poor fit for a virtualized system, particularly in 2010-2011
  - Recent developments in network interconnect driver virtualization and SR-IOV may be changing this
  - Application slowdowns were extreme
- Serial workloads were not substantially impacted in terms of performance
  - This led us to focus on bioinformatics workloads in general
  - Good performance for these tasks
- Our initial evaluation strategy was application centric
  - Tracking the application throughput bottleneck I described earlier
  - This proved to be a mistake

Inter Virtual Machine Networking

Provides an Ethernet L2/L3 interface to virtual machines

- Local connectivity between instances in the same security domain
- L3 security rules
- Ability to plumb in L2 networks
- *Ethernet* centric model
- Regular IPoIB won't work
- EoIB seemed promising, but drivers posed a problem

Inter VM Futures

Ethernet over IPoIB is designed to provide standard ethernet L2/L3 interfaces

- Drivers in beta
- Just got it working this morning (on host, not guest yet)
- Still working on performance
  - 15 Gbit/s aggregate over several TCP streams
  - 25 us latency (compared with 30 us over gige)
- And we need to see how bridging does

SR-IOV looks promising

- Drivers now available in beta
- But it likely won't fit the abstractions needed for non-HPC applications

Wide Area Networking

Provides VM instances with connectivity out of the Openstack deployment

- Implements L3 services
- Often requires address translation (from RFC1918 addresses to routable addresses)
- Implemented in software
  - With all of the performance problems that entails

Infiniband can definitely help

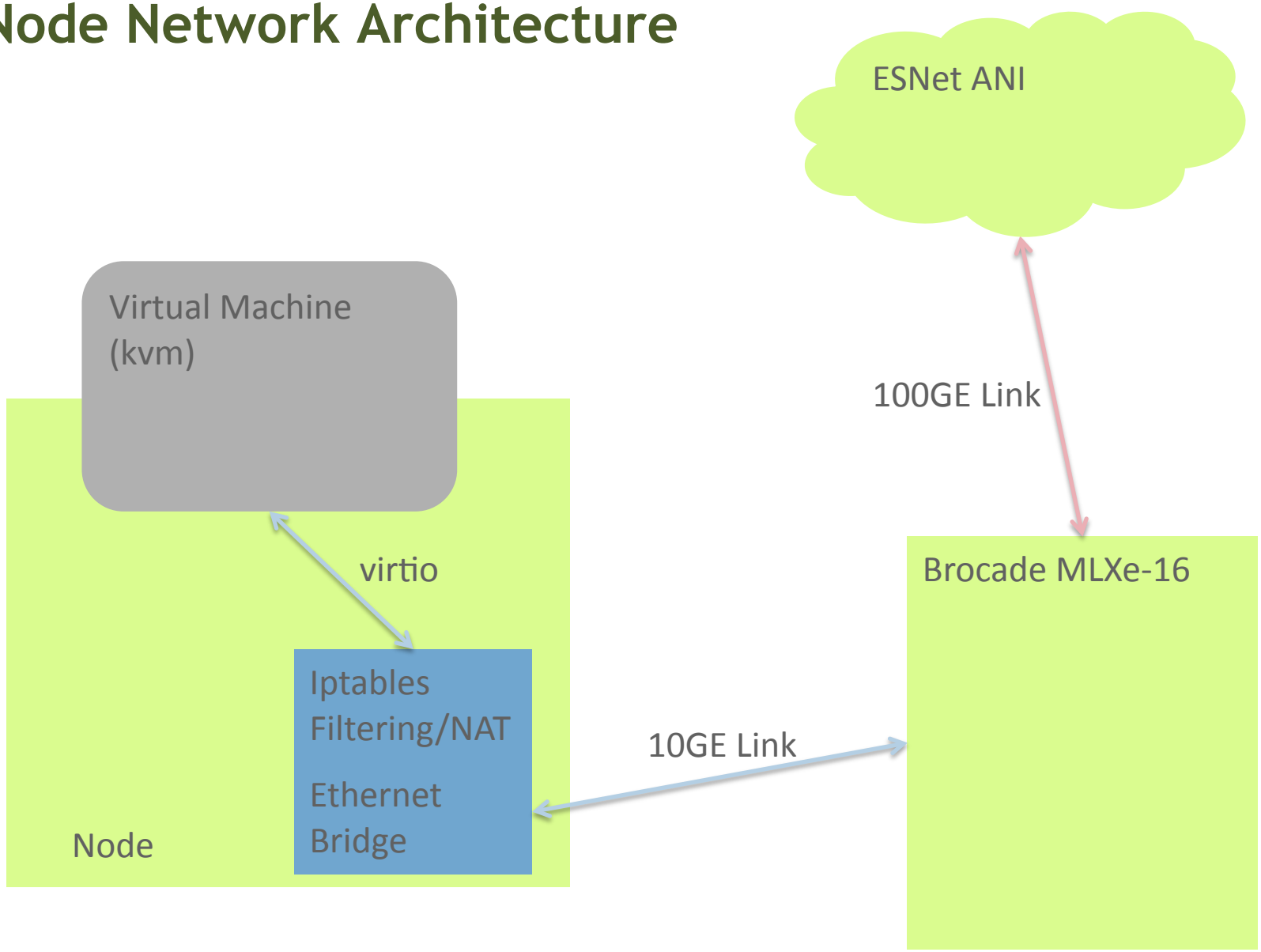
- Far better link rates into VMs (via eIPoIB, hopefully)
- But will need to integrate with upstream networks
  - Potentially Openflow
- Gateway features of SwitchX could potentially help as well

Wide Area Network Performance Expedition

- Goal: To determine the limits of Openstack infrastructure for wide area network transfers
  - Want small numbers of large flows as opposed to large numbers of slow flows
- Kbase will eventually need to support movement of 10-100's of TB of data per day
- Built a new Essex test deployment
  - 15 compute nodes, with 1x10GE link each
  - Had 15 more in reserve
  - Expected to need 20 nodes
  - KVM hypervisor
- Used FlatManager network setup
  - Multi-host configuration
  - Each hypervisor ran ethernet bridging and ip firewalling for its guest(s)
- Nodes connected to the DOE ESNet Advanced Networking Initiative

ESNet Advanced Networking Infrastructure

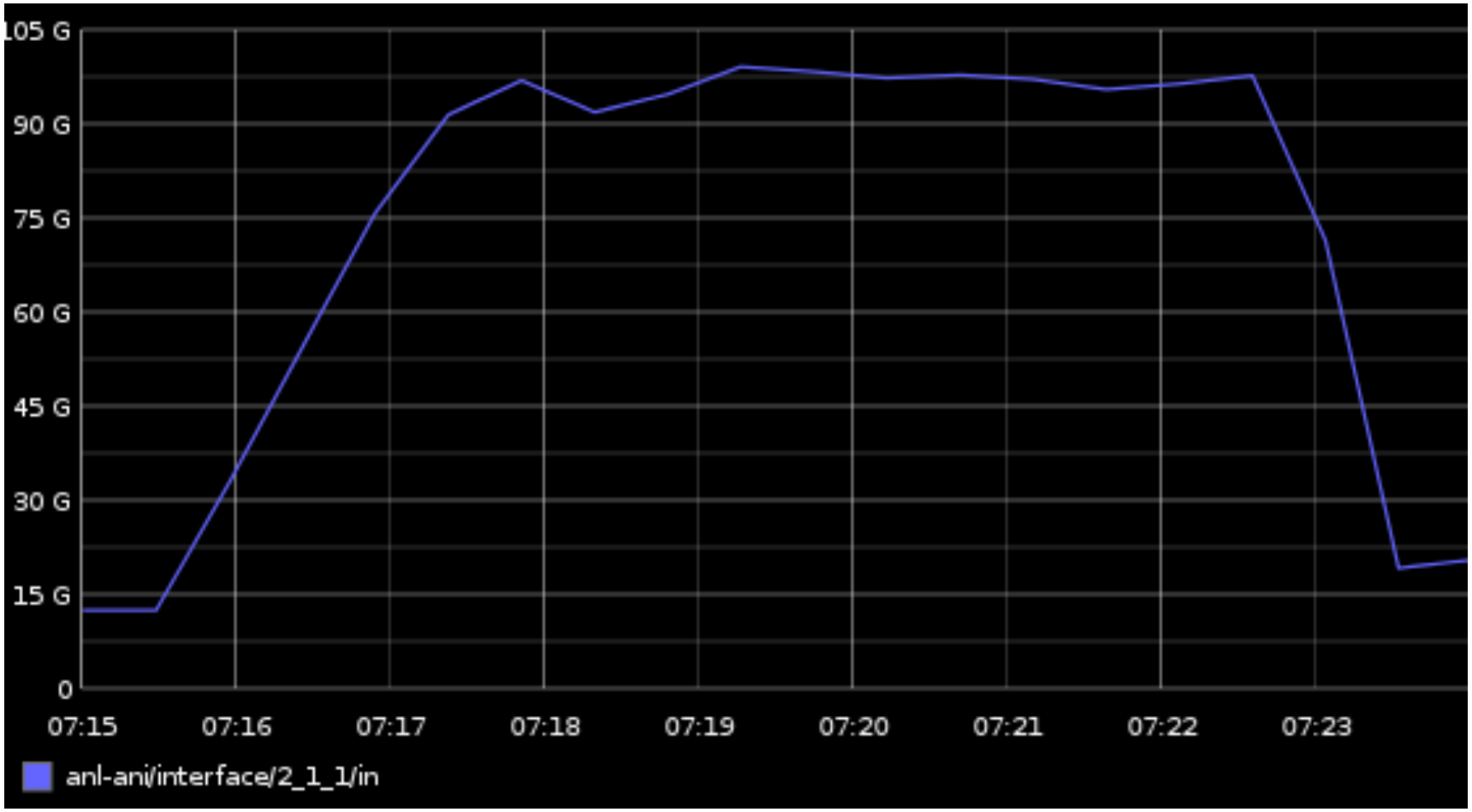
VM/Node Network Architecture



Setup and Tuning

- Standard instance type
  - 8 vcpus
  - 4 vnics bridged to the same 10GE ethernet
  - virtio
- Standard tuning for wide area high bandwidth transfers
  - Jumbo frames (9K MTU)
  - Increased TX queue length on the hypervisor
  - Buffer sizes on the guest
  - 32-64 MB window size on the guest
  - Fasterdata.es.net rocks!
- Remote data sinks
  - 3 nodes with 4x10GE
  - No virtualization
- Settled on 10 VMs for testing
  - 4 TCP flows each (ANL -> LBL)
  - Memory to memory

Network Performance Results



Expedition Results and Comments

- 95 gigabit consistently
  - 98 peak!
  - ~12 GB/s across 50 ms latency!
  - Only 10 nodes, with 1 vm/node
- Single node performance was way higher than we expected
  - CPU utilization even suggests we could handle more bandwidth (5-10 more?)
  - Might be able to improve more with EoIB or SR-IOV
- Single stream performance was worse than native
  - Topped out at 3.5-4 gigabits
- Exotic tuning wasn't really required
- Openstack performed beautifully
  - Was able to cleanly configure this networking setup
  - All of the APIs are usable in their intended ways
  - No duct tape involved!

Conclusions

- It is early days in high performance clouds
  - The software is not quite ready
  - But the flexibility of cloud control planes is clearly valuable, even on technical applications
- That said, we think that we'll be able to run HPC applications in the next few months
  - SR-IOV will enable low latency apps; initial reports from other sites are good
  - EoIPoIB will provide better connectivity between VMs for non-HPC applications
- One key issue will be how to marry IB networks with wide area SDN
  - Openflow, etc
- Also need to figure out how multi-tenancy should work
  - Pkeys are a start