JÜLICH
FORSCHUNGSZENTRUM

Supporting Lustre Community Testing

September, 24th 2012 | Frank Heckes, JSC, FZ-Jülich

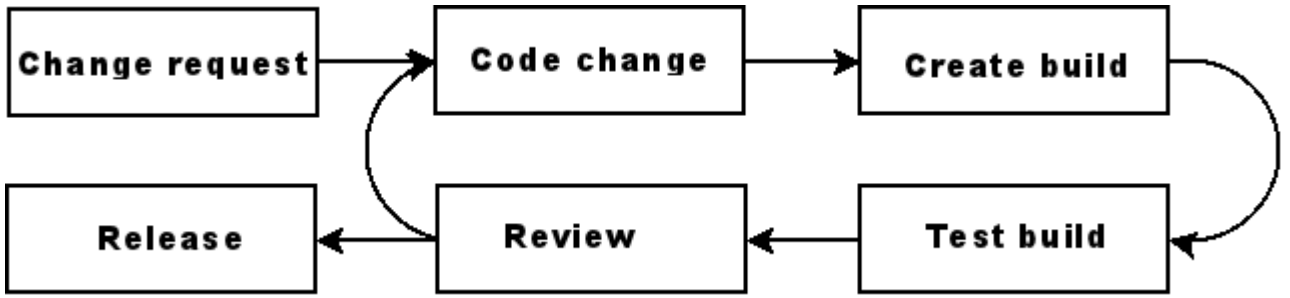
JÜLICH
FORSCHUNGSZENTRUM

Outline

- ***Overview***
- ***Test cluster description and experience***
- ***Improvements***

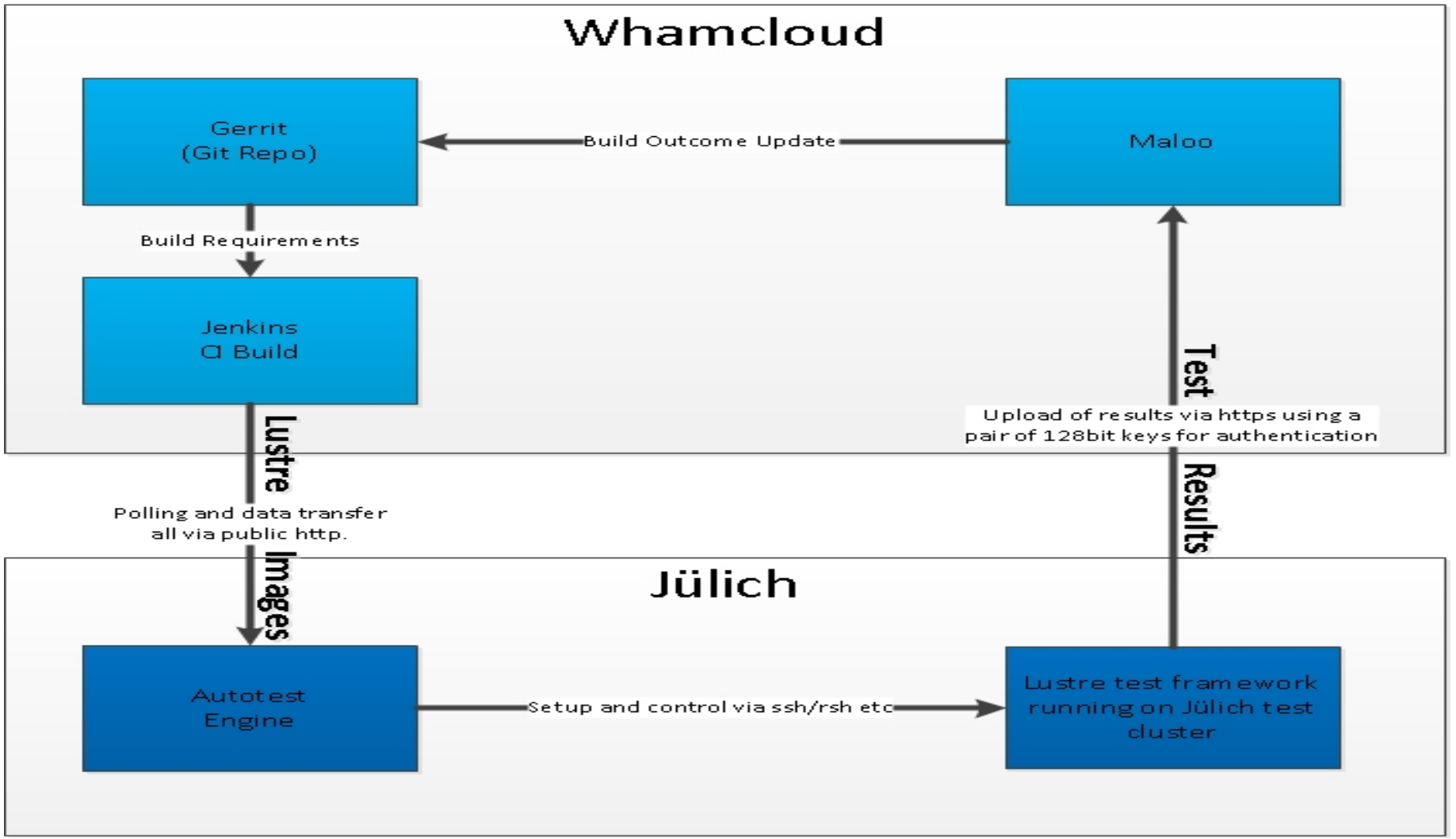
Lustre Community Test Cluster

- ***Whamcloud established new development model for Lustre:***



- ***Software Test performed with automated test frame-work***
- ***Chance to contribute to Lustre development by providing test infrastructure***

Lustre Community Test Cluster (logical view)



By courtesy of Chris Gearing, Whamcloud

Lustre Community Test Cluster

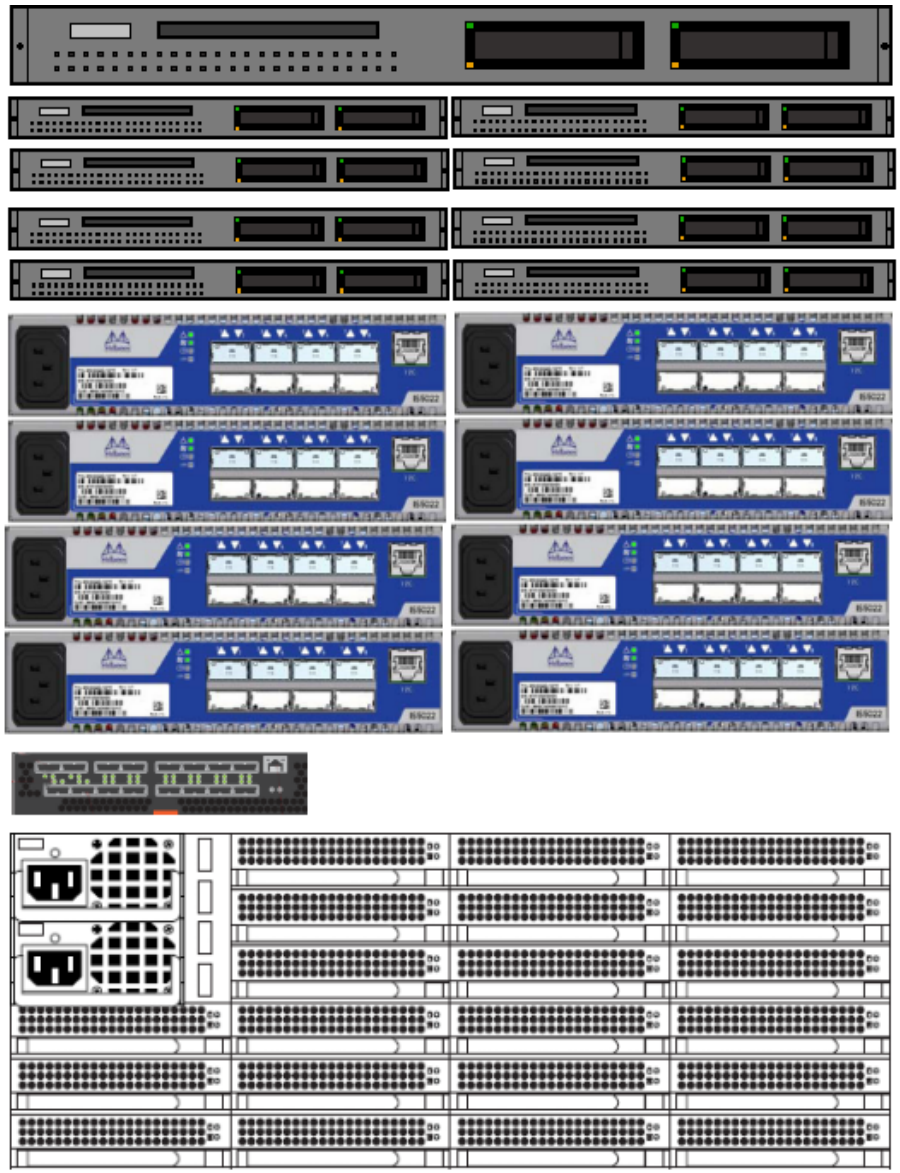
- ***Compiled requirement list together with Lustre engineering***
  - *At least 2 OSS, MDS nodes, (failover test)*
  - *At least 2 client nodes*
  - *Dedicated Infiniband fabric*
  - *Enough disk capacity for large LUN testing*
  - *CPU with virtualization capabilities*
  - *Sufficient memory*

JÜLICH
FORSCHUNGSZENTRUM

Lustre Community Test Cluster

- ***Hardware (SGI)***
  - *1 x Head Node*
  - *4 x Server Nodes, 4 x Client Nodes*
  - *8 x Mellanox Switches 8 Port Switches (small Full FAT tree)*
  - *SAS Switch*
  - *2 x JBOD (9 x 3 TB, 25 x 2 TB disks)*
- ***Configuration***
  - *Storage allocation via SAS zoning*
  - *Different autotest configuration map to different test cases (e.g.: large LUN testing, failover testing,..)*

Lustre Community Test Cluster



head node — autotest master, steer installation, send results

client nodes

OSS nodes

MDS nodes

Lustre Test Nodes

IB Switches

Lustre Network
JSC IB Test Network

SAS switch

JBOD

Lustre
Backend
Storage

JÜLICH
FORSCHUNGSZENTRUM

Lustre Community Test Cluster

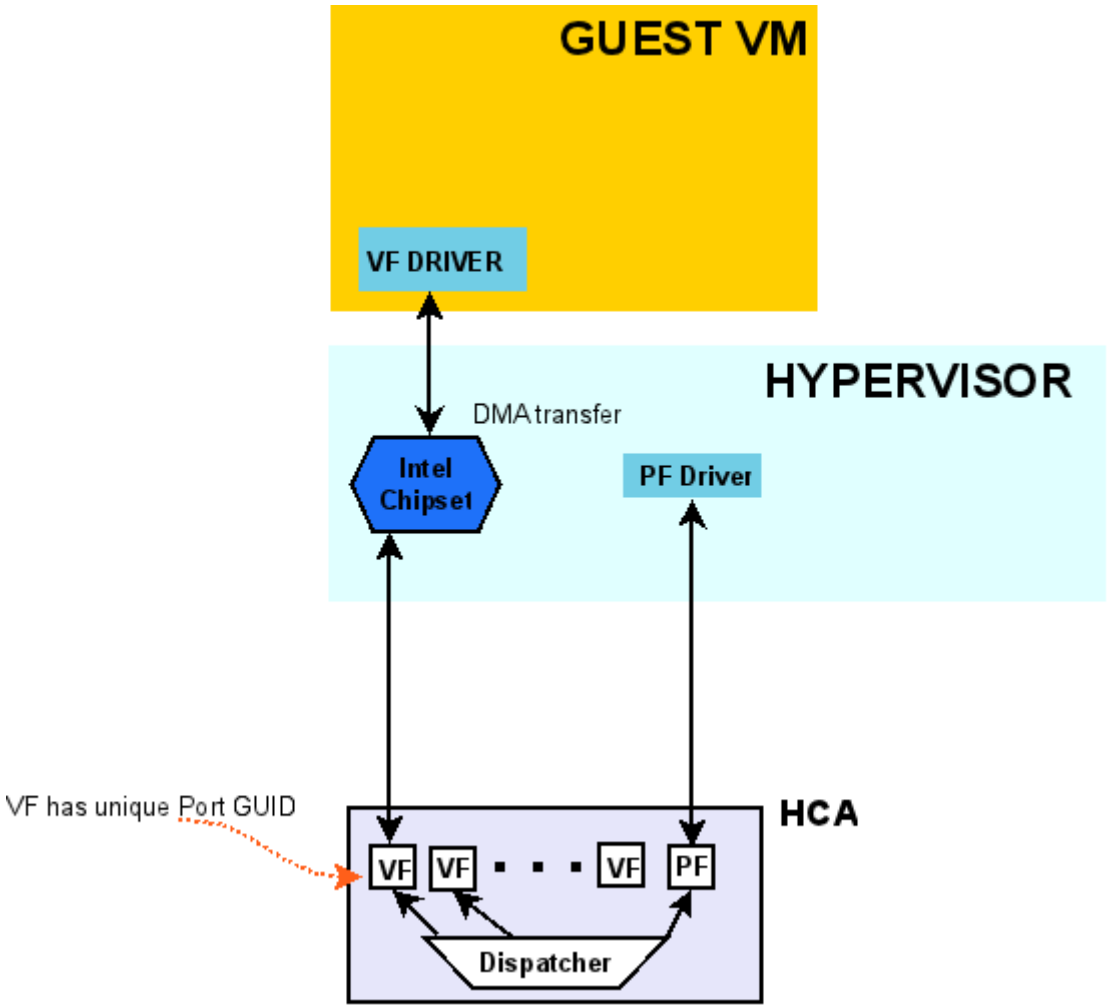
- ***Project Details***
  - *Small full FAT tree Infiniband fabric*

    → JSC tool development & IB hands on
  - *1'st Community Installation*
    - *Chris Gearing, Whamcloud (onside)*
    - *2 days for installation: Cobbler + autotest + infrastructure*
    - *Framework RHEL (CentOS) centric*
    - *Used in hundreds of test cases for Release 2.2*
    - *(see http://maloo.whamcloud.com → test result → user 'Juelich autotest')*
- ***Experience***
  - *Low administrative effort (< 1 Hour / month; on the average)*

Lustre Community Test Cluster

- ***Drawback***
  - *Test small test coverage due to small number of components*
  - *Need for Improvement*
- ***Mellanox provided (alpha version) Driver to virtualize HCA***
- ***Practical consequences:***
  - *Test coverage can be increased (factor 4-8)*
  - *Decomposition of test sets → Testing in parallel → reduced test time*
  - *Installation time can be reduced*
  - *Convenient way to use virtual machines for kernel debugging*
  - *Avoid NUMA I/O, Storage virtualization, client check-pointing*

Lustre Community Test Cluster



- ***Technical aspects***
  - *PF driver makes VF available*
  - *Guest use VF for direct communication*
  - *VF visible as HCA in guest OS*
  - *Each VF communicate 'independent' from the other*
  - *Up to 256 VF*

JÜLICH
FORSCHUNGSZENTRUM

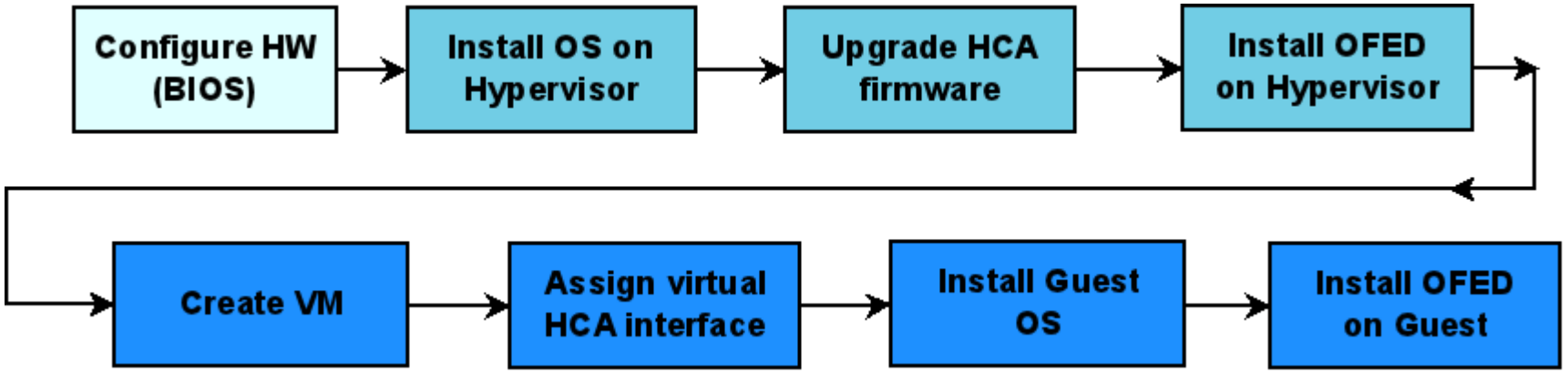
Lustre Community Test Cluster

- ***Hardware Requirements:***
  - *SR-IOV Support of mainboard*
  - *CPU Virtualization (Intel)*
  - *HCA: ConnectX2, ConnectX3 ASIC*
  - *Nice to have: IRQ remapping*
- ***Software Requirements***
  - *RHEL AS 6.2 (CentOS)*
  - *Kernel 2.6.32-220.13.1*
  - *KVM shipped with RHEL AS 6.2 (CentOS 6.2)*

Lustre Community Test Cluster

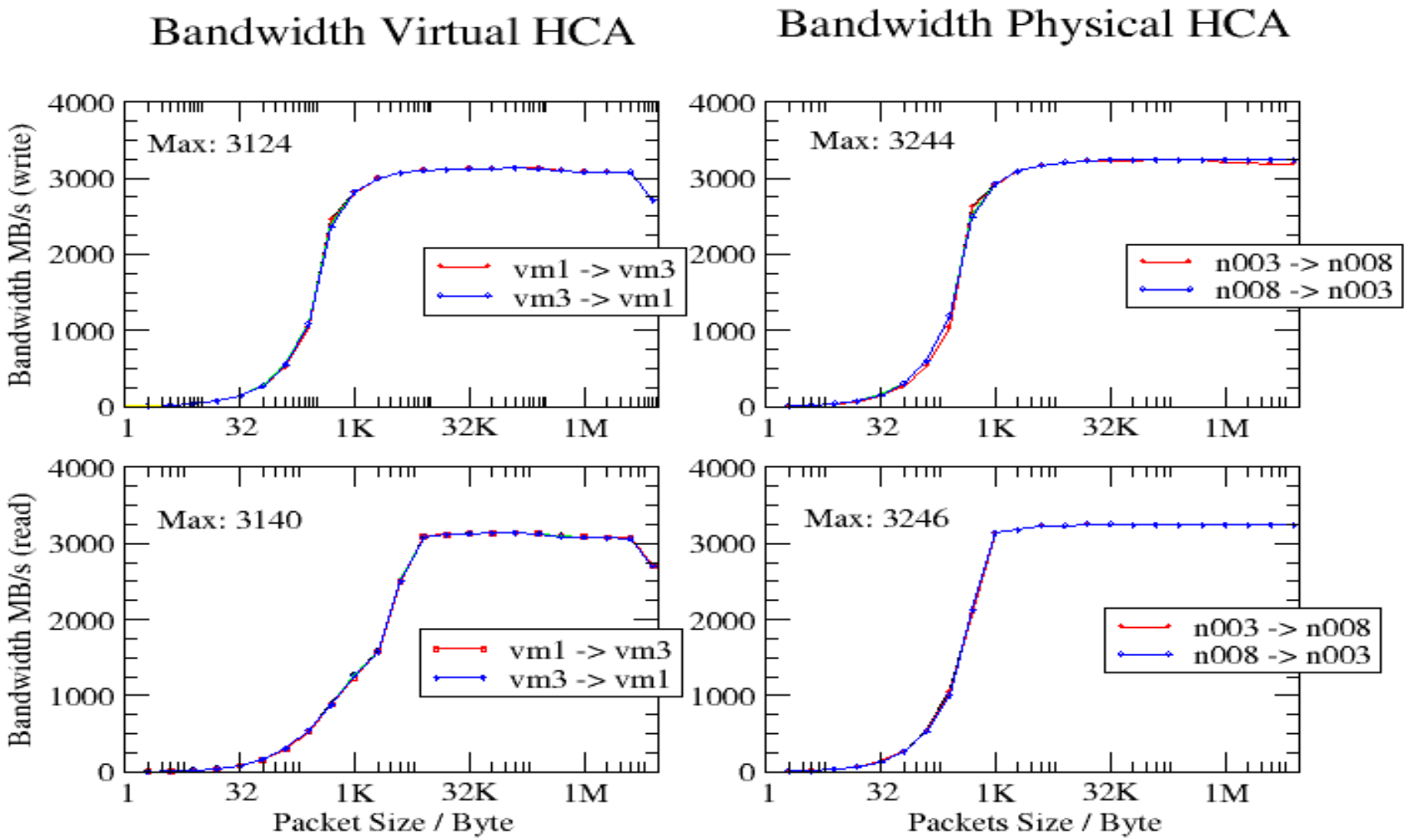
Software Changes

- ***HCA***
  - *FW (2.10.2000) has to be flashed*
- ***OFED***
  - *Need to run opensmd shipped with HCA (alpha) driver*
  - *(on hypervisor or network management node)*



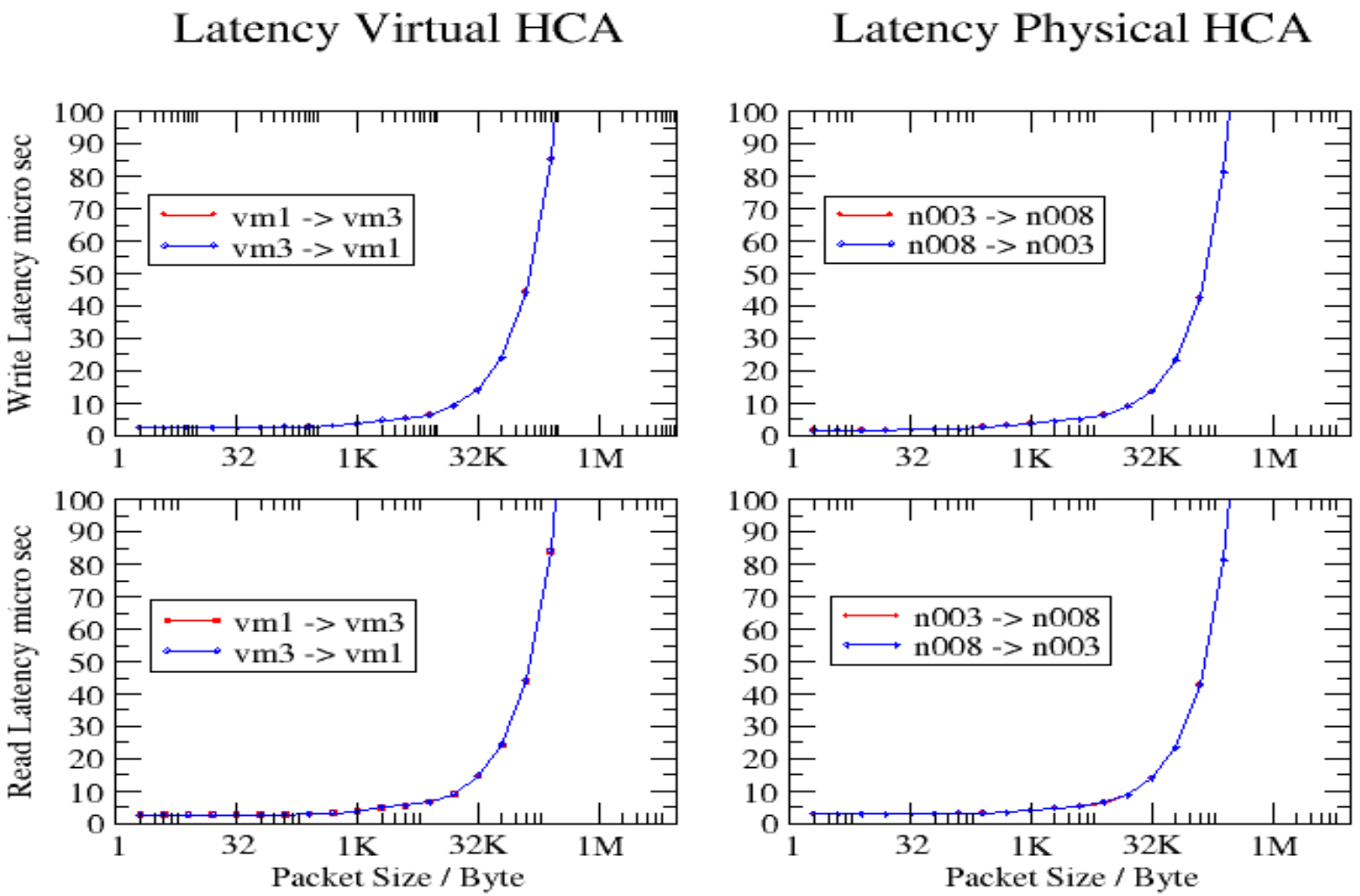
Lustre Community Test Cluster

- ***Bandwidth using ib_{read,write}_bw***



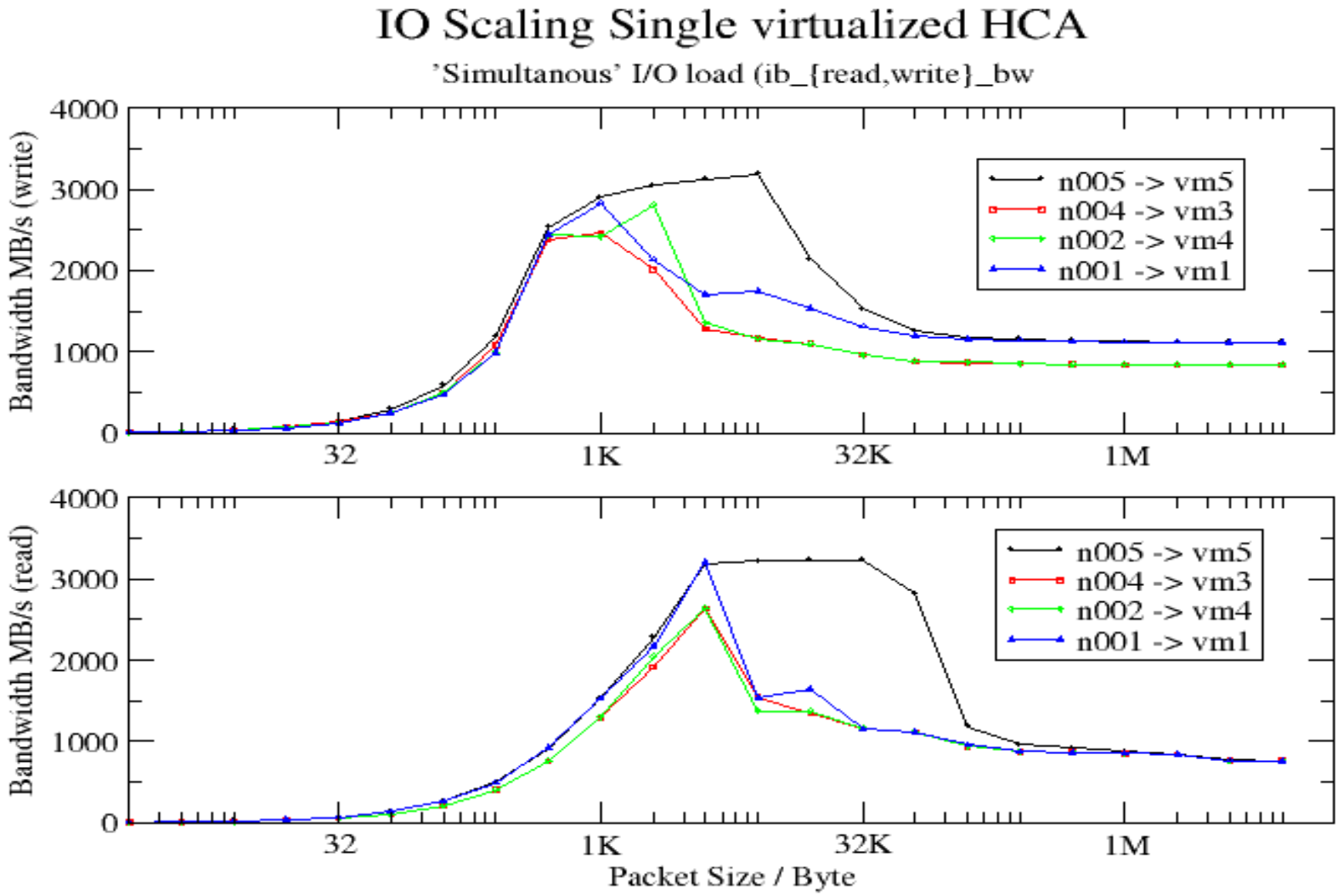
Lustre Community Test Cluster

- ***Latency (using ib_{read,write}_lat***



Lustre Community Test Cluster

- ***Simultaneous IO to 4 VF from 4 HCA***



JÜLICH
FORSCHUNGSZENTRUM

Lustre Community Test Cluster

- ***Todo***
  - *Solve missing 'support' for guest PXE boot for* `cobbler`
  - *'Framework' in autotest to handle resource allocation for different test scenarios*
    - *Only physical nodes*
    - *Only virtualized nodes*
    - *Mixture between physical / virtualized nodes*
  - *Get official Mellanox offical HCA firmware + OFED*
  - *Compile Lustre against Mellanox OFED*

JÜLICH
FORSCHUNGSZENTRUM

Lustre Community Test Cluster

- ***References***
  - *Lustre Test Results*
  - *http://maloo.whamcloud.com*
  - *SR-IOV*
    - *Specification:*
      http://www.pcisig.com/specifications/iov/
    - *Intel Video:*
      http://communities.intel.com/community/wired/blog/2010/09/07/sr-iov-explained
  - *KVM Bug*
    https://bugzilla.redhat.com/show_bug.cgi?id=715555

JÜLICH
FORSCHUNGSZENTRUM

Lustre Community Test Cluster

- ***Acknowledgment***

Chris Gearing (Whamcloud)

Thomas Husemann (Mellanox)
→ Providing SR-IOV HCA driver

JÜLICH
FORSCHUNGSZENTRUM

Questions ?