Infiniband stack integration

Marc Mendez
HPC Systems Architect

Agenda (1/2)

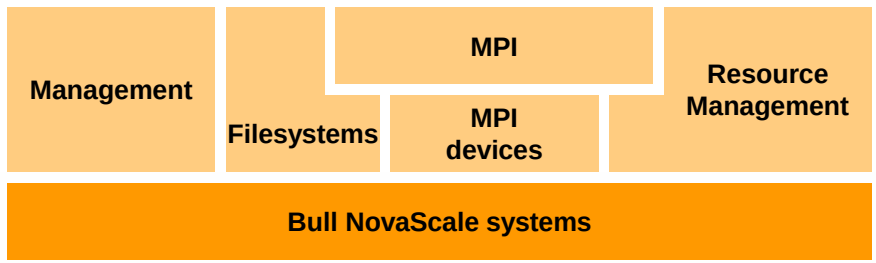
- Bull HPC Systems
  - IA-64 nodes
  - Interconnect networks
- Bull Advanced Server distribution
  - Cluster management
  - Batch and resource management
  - File-systems
  - MPI libraries

Agenda (2/2)

- Infiniband impact on management framework
  - Impacted components
  - Administration
  - Monitoring
- Infiniband impact on compute units
  - Resource management issues
  - MPI libraries for ISV
  - MPI Bull 2
- And as a result …

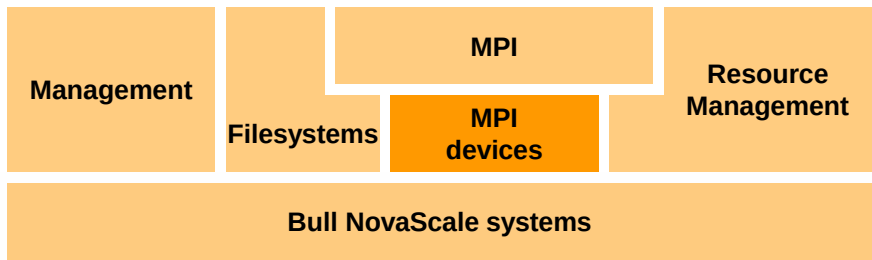
Bull HPC systems

- IA-64 based from 2 to 32 sockets
- Design includes NUMA and big I/Os
- NovaScale 5xxx supports PCI-X and, in a near future, PCI-e
- NovaScale 3xxx supports PCI-e natively and PCI-X for compatibility



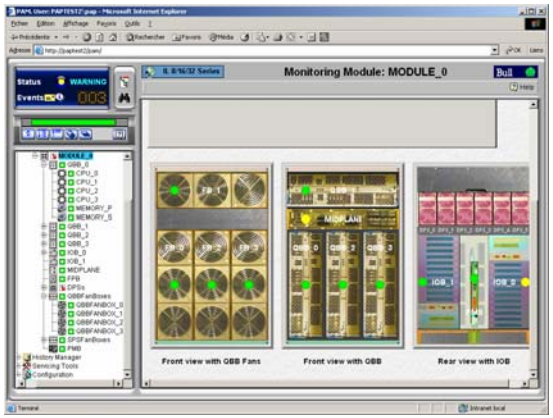
Interconnect networks

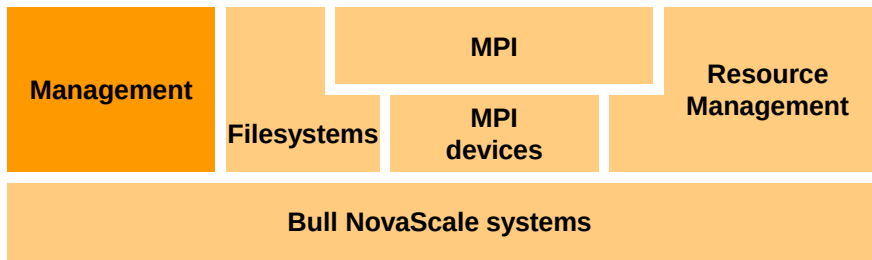
- *Quadrics* is our historical partner
  - Elan4/PCI-X gives 2.5 µs and 920 MB/s
  - Quadrics proprietary software
- We will develop *Infiniband* in the same way
  - 4x DDR/PCI-e 8x mainly, but …
  - Why not using SDR/PCI-e ?
  - We are awaiting better latencies and more bandwidth !
  - OpenFabrics and Slurm



BAS: cluster management

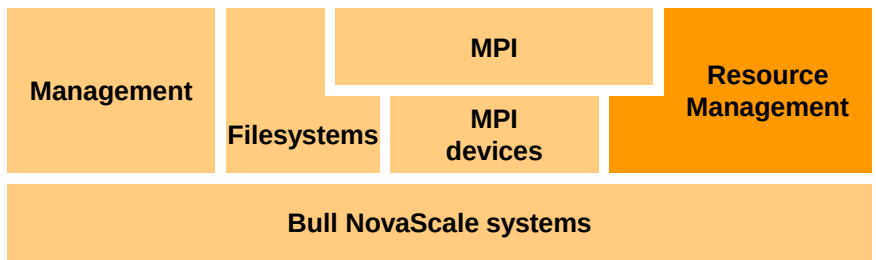
- One single interface to manage the cluster
- Open-Source tools have been extended/merged/integrated for consistency
- Features: cluster administration, performance/HW/SW monitoring, deployment, …
- Ability to interface 3rd party SW or HW with shell scripts, SNMP and APIs





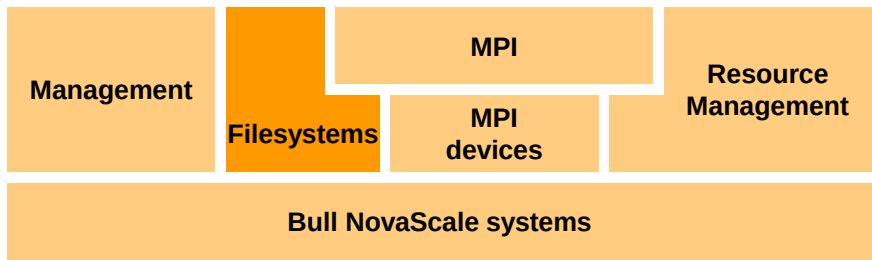
BAS: batch and resource management

- Users' computing resource entry point
- LSF often required, Torque(/MAUI) available
- Until now, Quadrics RMS was used
- RMS may manage Infiniband networks but …
  - Is not free
  - Tons of features but lack of modularity
  - Source code is not available



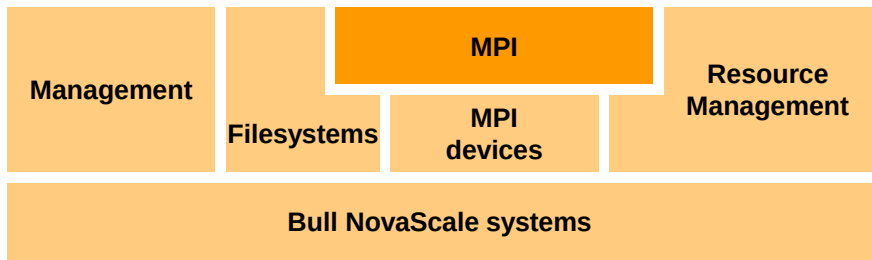
BAS: filesystems

- **Local filesystem: EXT3 or XFS**
- **Parallel filesystem: Lustre**
  - **Chosen for scalability, robustness and performances**
  - **Big partnership with CFS to integrate and optimize Lustre**
  - **DDN systems fits HPC performance requirements**
  - **Exclusive administration/configuration tools**



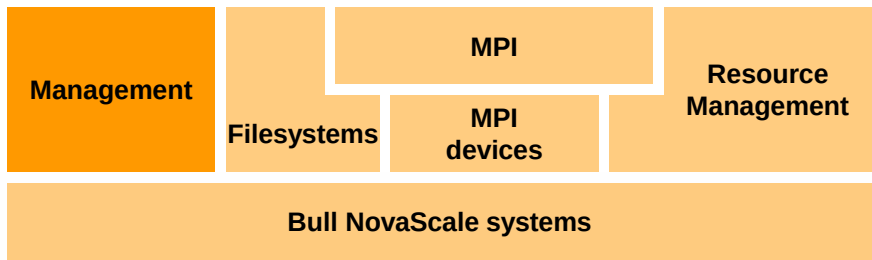
BAS: MPI libraries

- **Two requirements:**
  - **MPI Bull 2 for performances**
    - ***Postal* optimizes intra-node latencies and bandwidth**
    - **Dynamic framework for ADI3/CH3-OSU devices fast and easy integration**
  - ***MPI* for ISVs**
    - **mpich-ethernet provided**
    - **May come with HP/MPI, Intel MPI, …**



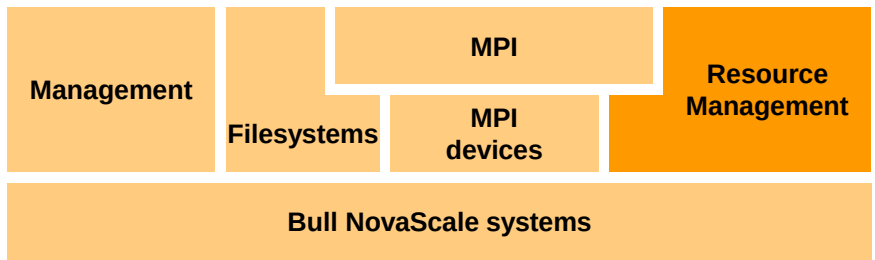
Infiniband impact on management framework

SW monitoring	SM, Resource Manager …
HW monitoring	Infiniband devices
Performance monitoring	Infiniband counters
Administration	Network topology, Resource Manager …
Deployment	



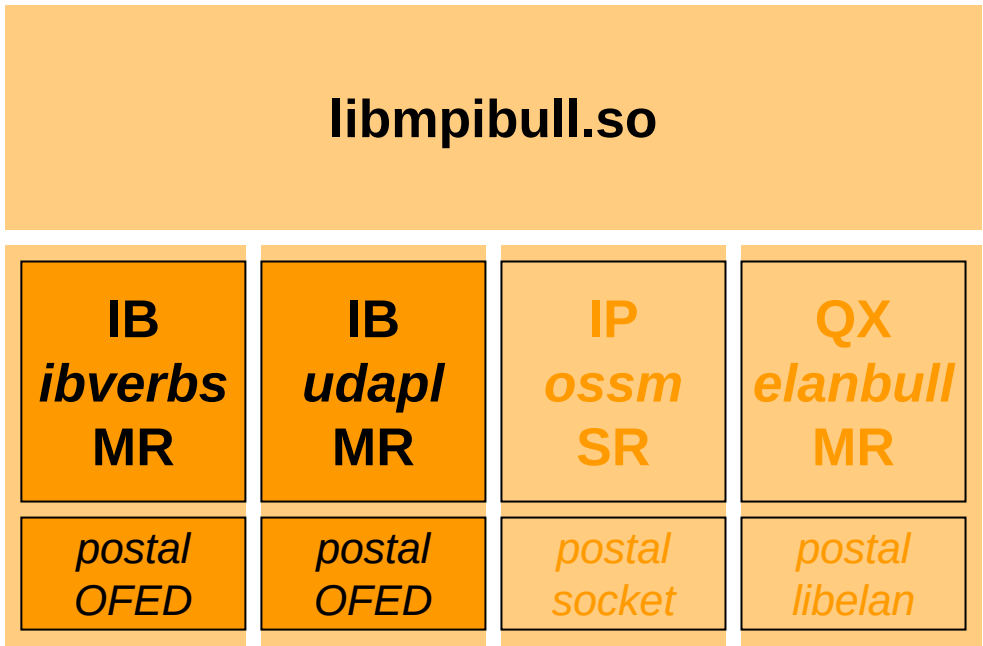
Infiniband impact on compute units (1/4)

- **Using Slurm, we have to:**
  - **Integrate CPUSET for NUMA systems efficiency**
  - **Enhance accounting information**
  - **Provide Platform/LSF and Torque coupling interfaces**
  - **Improve PMI scalability for MPI Bull 2**
- **However, Slurm is:**
  - **Simple, modular, open-source, and**
  - **May replace RMS even when using a Quadrics network.**



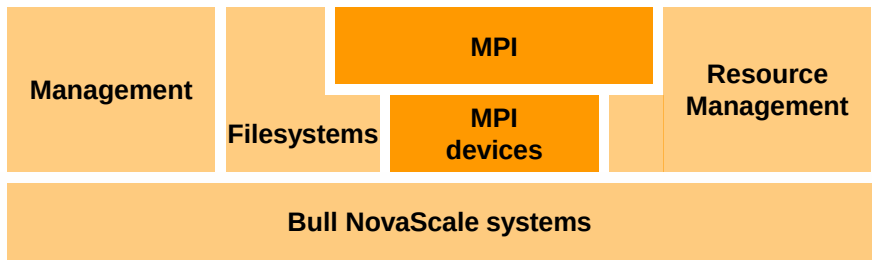
Infiniband impact on compute units (2/4)

- **MPI Bull 2: compile once, run every IA-64 !**



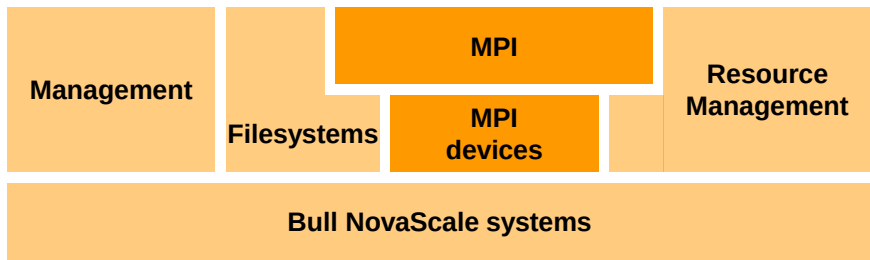
MPI Bull 2 core library (dynamic and static)

MPI Bull 2 device drivers libraries (dynamic)



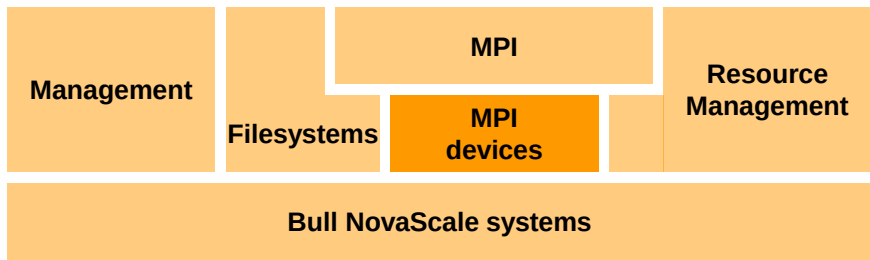
Infiniband impact on compute units (3/4)

- *Postal* is designed to reach the best NUMA performances:
  - SMD: sub-microseconds latencies (0.55µs with Madison 1.6GHz) using lock-free mechanisms
  - MDM: full-memory bandwidth available to applications using zero-copy (4.2GB/s on NS3xxx) and (almost) lock-free design. Provides optimized one-sided routines.
- *Postal* is network ready:
  - callbacks to customize ANY_SOURCE receive requests



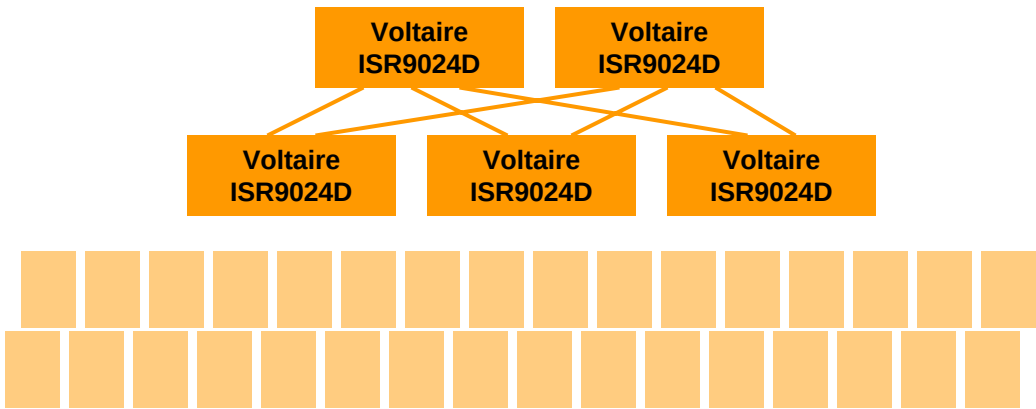
Infiniband impact on compute units (4/4)

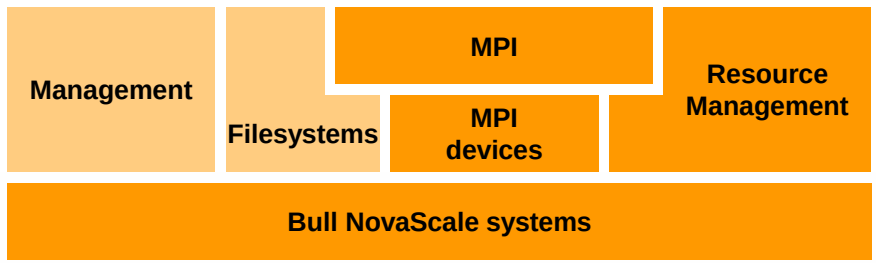
- MPI “device drivers” may be integrated in a coffee-break time (or so) if ADI3 or CH3/OSU compatible !
- Infiniband *ibmr_gen2* “device driver” has been stolen from MVAPICH-2 and will be the target for optimal performance
- *ibmr_gen2* “device driver” may be wildly tuned (and is documented). We imagine in a mid-term future to integrate adaptative algorithms to enhance performances.



And as a result … (1/3)

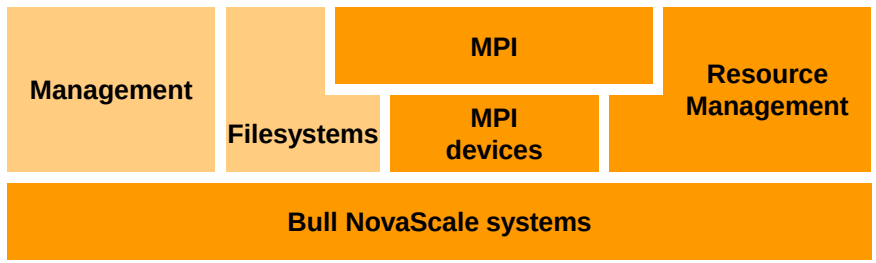
- We built a development/testing cluster:
  - 32x 8-core 32GB compute nodes (~1.6 Tflops w/ Montecito 1.6GHz/18MB)
  - Voltaire 4x DDR two switch-levels non-blocking network
  - MemFree single port HCAs





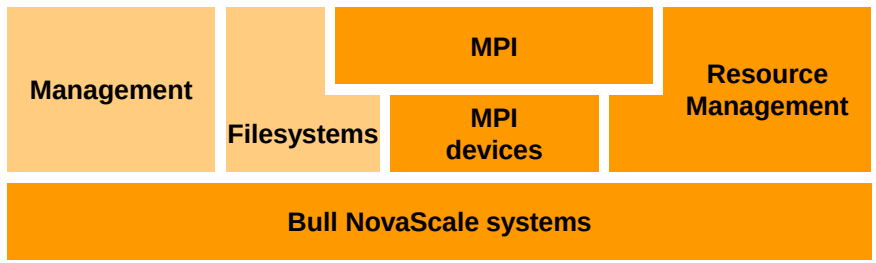
And as a result … (2/3)

- Switch configuration took 1 hour (because I was scrutinizing around the interface ;))
- Network debugging was fast (thanks to Voltaire support): no fault !!!
- OFED 1.0 RC4 integration to 2.6.12 Bull kernel
- Infiniband network and stack was stable however !
- A bug appeared in Slurm PMI module over 96 cores (we then used MPD as a fallback)



And as a result … (3/3)

- Resource manager needs improvements !
- Bandwidth: > 1.3 GB/s unidirectionnal
- Latency: ~ 6 µs MaxPingPong
- Every benchmark was run successfully:
  - HPCC
  - IMB
  - ESX …
- Switch to up-to-date 2.6.16 kernel is ongoing and should deliver much better results.



Architect of an Open World™