



High Productivity MPI – Grid, Multi-cluster and Embedded Systems Extensions

Presented by

Dr. Anthony Skjellum

Chief Software Architect

Verari systems Software, Inc

September 30, 2004

HPEC Workshop

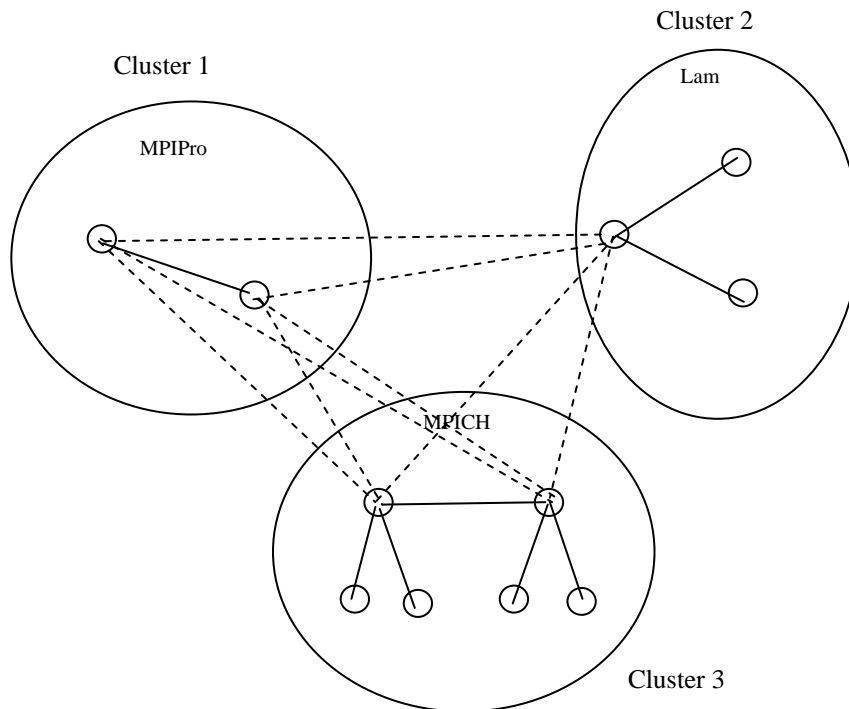
IMPI

- **IMPI – Interoperable Message Passing Interface**
- **Developed and Proposed by NIST**
- **Standard for inter-operation of multiple**
 - **Implementations (IMPI, IMPI-2)**
 - **Architectures**
 - **Networks**
 - **Operating Systems**

Client, Host Concept

- **MPI processes spread across multiple clients**
- **Clients represent MPI processes belonging to a single implementation**
- **Hosts represent gateways for processes of Clients**
- **IMPI Application may have two or more clients**
- **Client may have one or more hosts**
- **Hosts serve as gateways for one or more MPI processes**

Typical Scenario – Multi-vendor MPI



- **3 Clients (Each cluster make one client)**
- **Client 1**
 - **2 hosts, 2 MPI processes**
- **Client 2**
 - **1 host, 3 MPI processes**
- **Client 3**
 - **2 host, 6 MPI processes**

MPI/Pro 1.7.0

- **MPI/Pro 1.7.0 provides first complete implementation of IMPI**
- **Enables Interoperation between**
 - **Windows, Linux and Mac OSX operating systems**
 - **32-bit and 64-bit architectures**
 - **TCP, GM and VAPI Networks**
 - **Any combination of all the above**

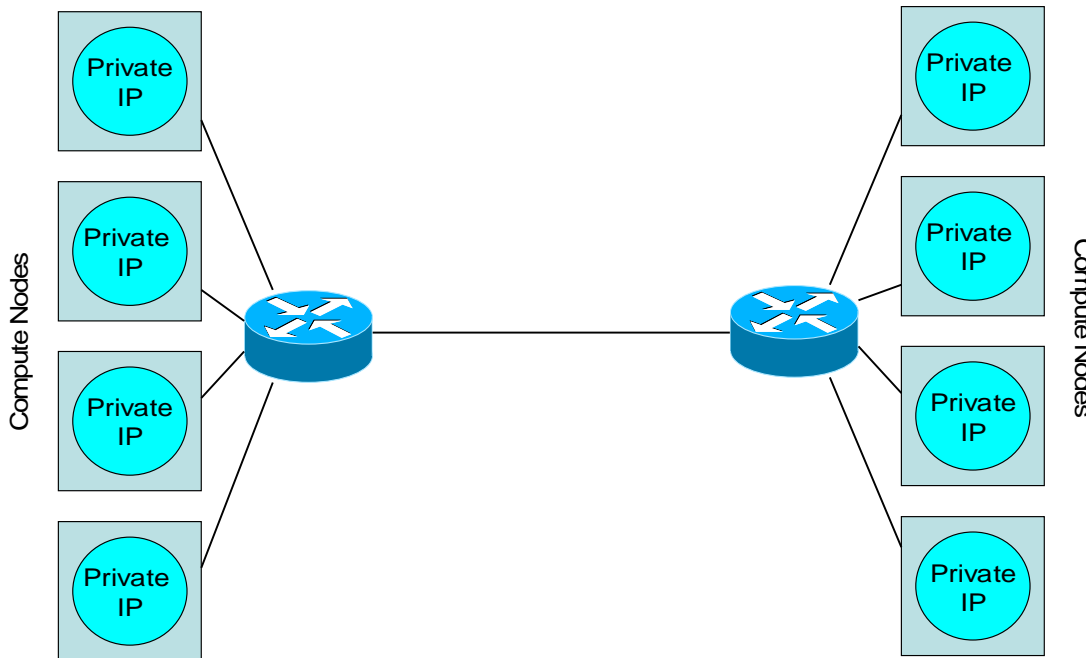
Extensions

- **IMPI does not address issues such as**
 - **Private IP Addresses**
 - **Owner domains**
 - **Faults**
 - **Interacting Dynamic Groups**
- **Above issues play vital role in Grid**
- **Verari proposed and implemented a novel method to address issue of Private IP Addresses**

Case Study

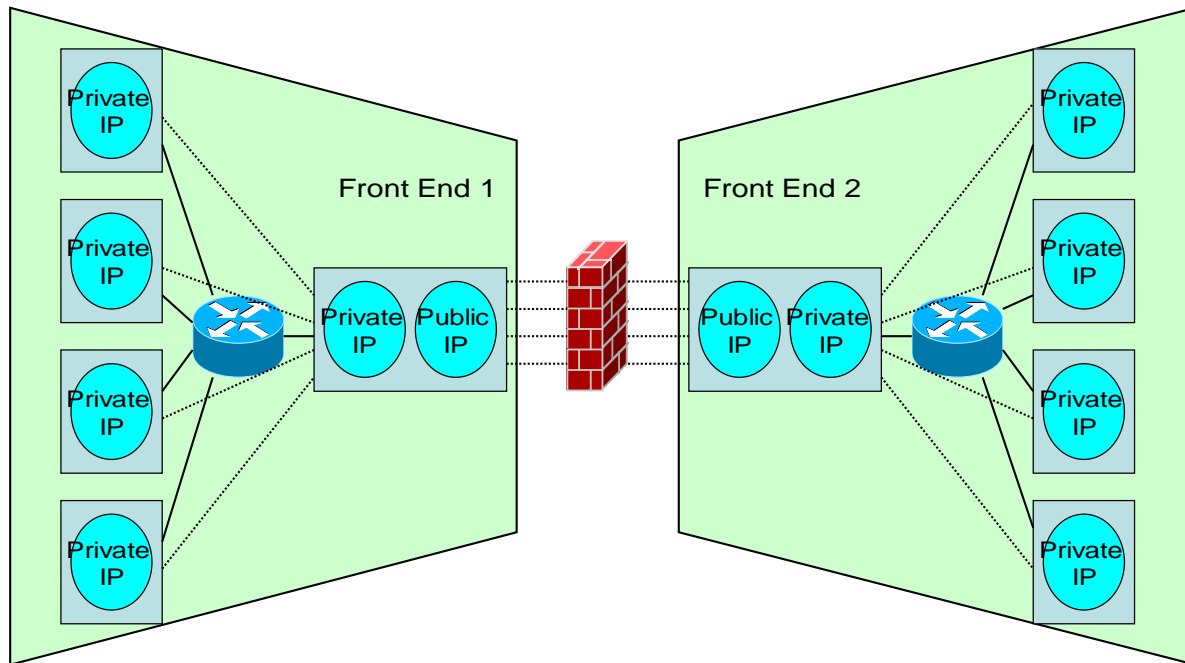
Private IP Enabled IMPI

Typical Cluster Setup



- **Compute Nodes have private IP addresses**
- **External communication through single head node or gateway**
- **Unsuitable for multi-cluster multi-site communication**

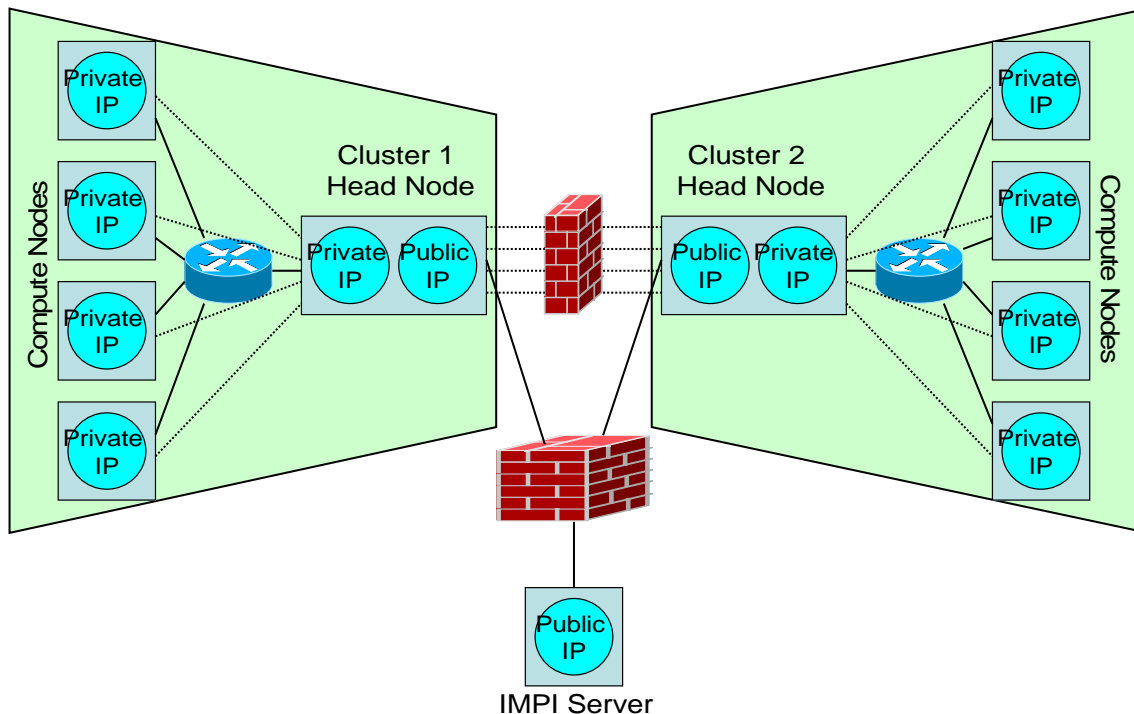
Network Address Translation (NAT)



- Firewalls block incoming connections

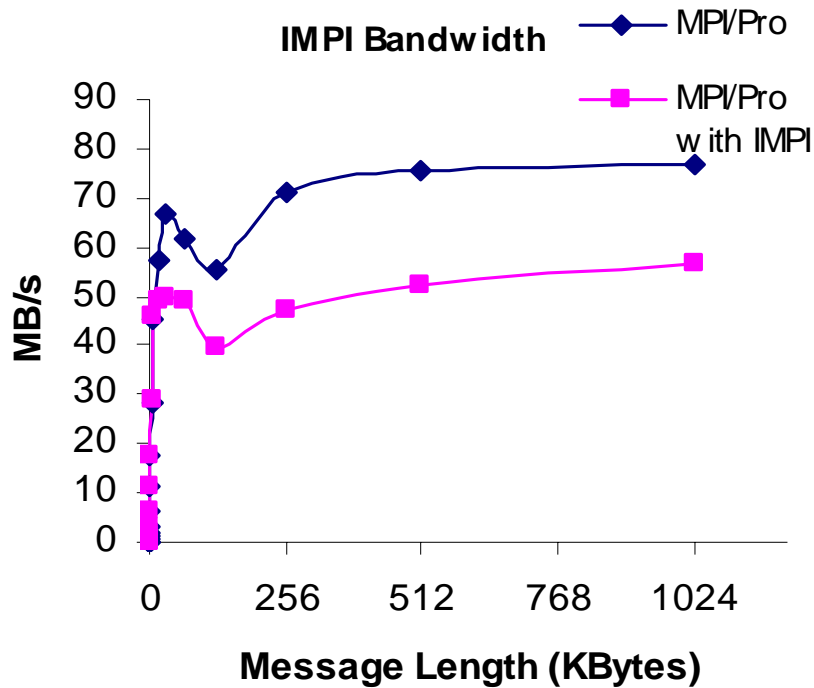
- NAT used to serve requests generated within the cluster

NAT-based IMPI



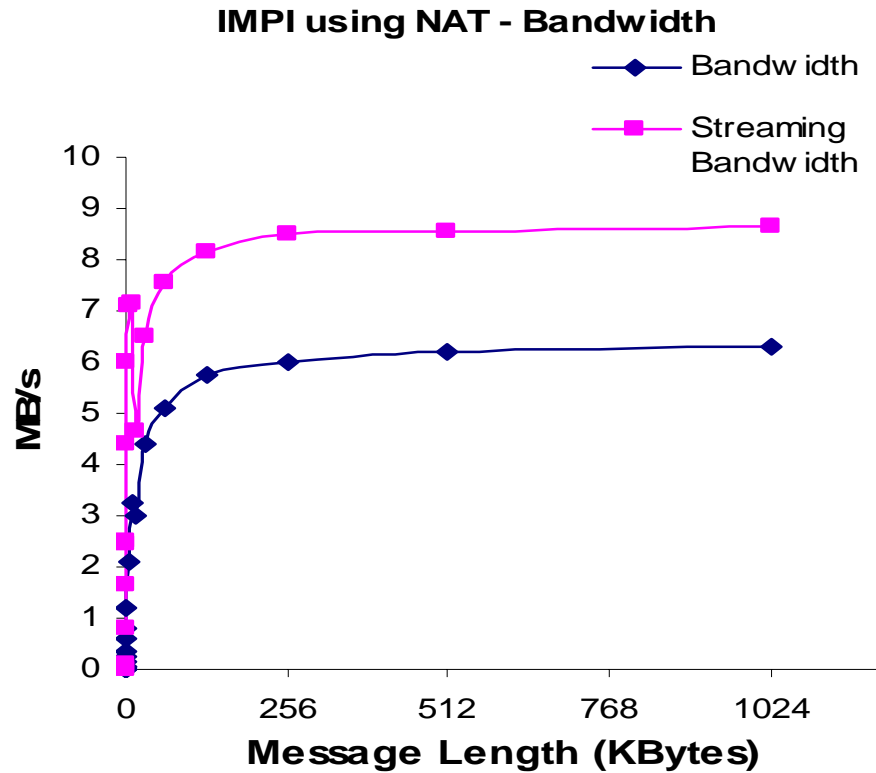
- Use NAT to generate dynamic mappings between head node and compute nodes
- Dissipate dynamic mapping info through IMPI server
- Release mapped ports on head node on completion of application

Performance



Configuration	Latency (us)
MPI/Pro without IMPI	142.45
MPI/Pro with IMPI	147.35

Performance



Proposed Extensions

- **IMPI extensions for MPI-2**
- **Open protocol-based initialization such as SOAP**
- **Adaptation to the Grid**
- **Reduce user involvement**
- **Optimize for performance**

References

- Velusamy, V *et al.* *Communication Strategies for Private-IP-Enabled Interoperable Message Passing across Grid Environments*, First International Workshop on Networks for Grid Applications, 2004.
- MPI Forum. *MPI: A Message-Passing Interface Standard*. 1994.
<http://www.mpi-forum.org/docs>.
- IMPI Steering Committee. *IMPI Draft Standard*. 1999.
<http://impi.nist.gov/IMPI/>.
- Gropp, W. *et al.* *A High-Performance Portable Implementation of the Message-Passing Interface*. *Journal of Parallel Computing*, 22, 1996, pp: 789-828.
- Skjellum, A. and McMahon, T. *Interoperability of Message-Passing Interface Implementations: a Position Paper*.
<http://www.verarisoft.com/publications/files/interop121897.pdf>