



## **ChaMPlon/Pro: The Complete MPI-2 for Massively Parallel Linux Clusters**

Rossen Dimitrov  
Verari Systems Software, Inc.

Linux Cluster Institute Conference  
Austin, May 19, 2004



### **Overview**

- Background
- Overview of ChaMPlon/Pro
- Supported Configurations
- Special Features
- Selected Performance Results
- Other Clustering Products from Verari
- Summary



## Verari

- MPI Software Technology recently merged with RackSaver and formed Verari Systems
- MPI Software Technology's software products are now available through Verari Systems
- Verari provides complete HPC solutions: high-density Xeon and Opteron blades, I/O, and complete cluster software stack and services



## Background

- ChaMPIon/Pro™ is a robust, scalable, high-performance, commercial MPI-2.1 implementation, with native MPI-IO (MercurIO)
- Builds on the success of MPI/Pro, our MPI-1 implementation, and experience from the DOE ASCI-PF program
- The first commercial MPI-2.1 available for Linux and MacOS

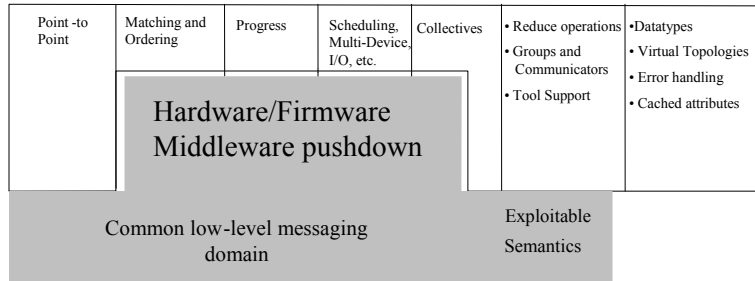
## Design Principles

- Emphasis on time to solution
- Balance of performance criteria (latency, bandwidth, CPU overhead) and resource utilization (memory, NIC)
- Performance predictability
- Works to retain system scalability for applications

## Design Principles, II

- Avoid artificial abstraction layers, used for portability of the MPI code itself
- Adding more software layers does not add to scalability, performance, or predictability [Brightwell's law]
- Multithreading and internal MPI concurrency

## Design Adaptability



## Design Taxonomy for MPI Implementations

- Classification criteria
  - Completion notification
    - Polling
    - Blocking
  - Message progress
    - Polling
    - Independent

<b>Polling notification</b> <b>Polling progress</b> MPICH, MPI/Pro's and ChaMPion/Pro's short protocol in polling mode	<b>Polling notification</b> <b>Independent progress</b> MPI/Pro's long protocol for some devices in polling mode
<b>Blocking notification</b> <b>Polling progress</b> (no practical use)	<b>Blocking notification</b> <b>Independent progress</b> MPI/Pro and ChaMPion/Pro in blocking mode



## Performance and Scalability

- Scaling to thousands of processors
- Thread safety (highest level)
- Optimized collective operations
- Optimized derived datatypes
- Efficient memory and NIC resource usage
- Multi-device support
- Topology awareness



## Functionality and Usability

- Integration with schedulers and resource managers
- Integration with debuggers and profilers
- Functionality controlled by tunable parameters
- Documentation
- Reflect user feedback



## Full MPI-2 Functionality

- MPI-1.2
- Parallel I/O
- One-sided communication
- Dynamic process management
- Extended collective operations
- Improved error handling
- Info object
- External interfaces (language bindings, profiling)



## Key Qualities

- Fully compliant, native MPI I/O
- Efficient one-sided communication
- Independent message progress
- Low CPU overhead
- Multi-device support
- Overlap of communication, computation, I/O
- Thread safety/awareness [MPI\_THREAD\_MULTIPLE]
- Works fully with OpenMP



## **MercurtIO: Fast MPI I/O**

- The MPI-IO Component of ChaMPlon/Pro
- Supports: NFS, ENFS, PVFS, GPFS, Lustre, Panasas
- Unique object oriented design w/ asynchronous I/O
- Performance optimizations for aggregation and non-contiguous file access with adaptive strategies



## **MercurtIO, II**

- Large File Support (64 bit file size)
- Split collective I/O uses asynchronous I/O
- Shared file pointer implemented with file lock (if supported by the file system) or one-sided communication lock
- Most advanced support for Parallel file I/O available for Linux clusters



## Important Capabilities for Production Computing

- C and C++ Language Bindings
- ISO FORTRAN 90 and FORTRAN 77
- PERUSE Support
- Improved error handling
- Extensive performance and correctness test suites
- Customizable



## Unique PERUSE Support

- Level of detail of MPI profiling not possible through PMPI
- Investigating hard performance and scalability issues.
- Can be used to study the behavior of the MPI middleware as well as the behavior of the hardware in greater detail
- Can be used to complement PMPI
- MPI profiling tools can utilize PERUSE to provide additional services for performance analysis



## Platform Support

- HPC clusters
- DOE ASCI platforms
- Embedded space



## ChaMPLion/Pro for Clusters

- OS
  - Linux
  - MacOS
- Processor architecture
  - Intel (ia32,ia64),
  - AMD (Opteron),
  - Apple (G5)



## Interconnect Support

Network	Interface
Myrinet	GM; GM-2
InfiniBand	VAPI
Quadrics (QsNet)	ELAN
100/1000 Ethernet	TCP
SMP	
*RACE/RACE++	DX

\*For embedded Mercury multi-computers



## File System Support

- Lustre
- Panasas
- PVFS
- NFS
- UFS



## Integration with Tools and Resource Managers

- Debuggers: TotalView; multiple gdb
- Profilers: SeeWithin/Pro; other PMPI
- Schedulers:
  - LSF
  - PBS
  - Cycles@Work
  - BPROC
  - SLURM



## ASCI Platform Support

- Sandia Cplant (HP/Compaq Alpha; Linux; Myrinet/Portals; ENFS; yod; yod2)
- LANL InfiniBand and GigE clusters (ia32; Linux; Panasas, LSF, BPROC)
- LLNL ASCI Blue & White IBM SP systems (PPC 603e & Power 3; IBM AIX; SP Switch/LAPI; GPFS; Gang/LL)
- LLNL Quadrics cluster (ia32; Linux; Quadrics/ELAN; Lustre File System; SLURM/pdsh)
- Tools: Etnus TotalView and Intel/Pallas Vampir



## Embedded Platforms

- Mercury PPC-based multi-computers
- RACE and RACE++ interconnect
- Upcoming support for RapidIO interconnect
- MC/OS (Mercury OS) 5.8 and 6.0
- Solaris, Windows, vxWorks hosts
- MPI-1 + full one-sided implementation
- 0-length message latency:  $< 4 \mu\text{s}$
- BW of 1 MB message  $> 245 \text{ MB/sec}$



## Special Features

- Scalability
- Choice of message completion notification
- Concurrent multi-device support

## Scalability

- Eliminate  $O(n)$ ,  $O(n^2)$
- Use  $O(1)$ ,  $O(\log n)$
- Scalability affects all aspects of design
  - Task startup
  - Initialization
  - Point-to-point communication protocols
  - Collective communication
  - Resource utilization (memory, NIC resources)

## Choice of Completion/Progress

- Polling mode
  - Polling completion; Polling progress
  - Lowest latency, wastes CPU cycles
  - Useful for apps with short messages
- Blocking mode (default)
  - Blocking completion; independent progress
  - Low CPU utilization; timely delivery of long messages; allows for overlapping

## Point-to-point in Different Modes

- Platform
  - Dell PE 1750
  - Dual Intel Xeon 3.06 GHz
  - RedHat 9
- Network 1
  - Myrinet PCIX-D
  - GM 2.0.2
- Network 2
  - GigE; tg3 drivers
  - Force10 switches

Device and Completion mode	Latency (microseconds) Msg size = 0	Bandwidth (MB/s) Msg size = 1 MB
Myrinet/GM-2 Polling	7.29	232.71
Myrinet/GM-2 Blocking	20.86	231.71
GigE/TCP Blocking	40.16	110.93

## HPL Linpack in Different Modes

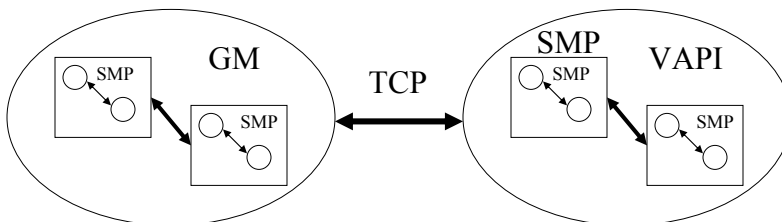
NB=104 N	GM Polling (GFLOPS)	GM Blocking (GFLOPS)
20000	164	160
40000	226	225
60000	253	251
80000	267	265
100000	277	281

Same platform; 64 nodes; 1 process per node

## Multi-Device Support

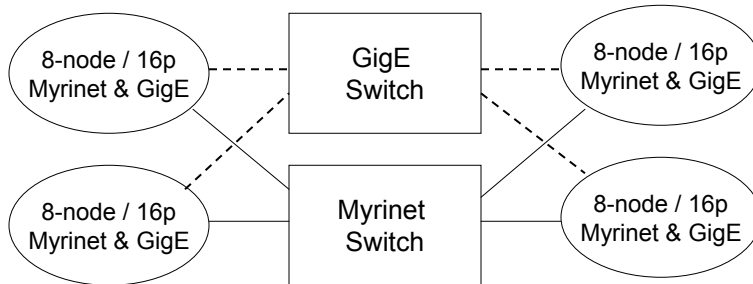
- Support for concurrent communication over multiple fabrics (e.g., SMP, TCP, and GM)
- Independent message progress on all devices
- Removes performance and scalability inter-dependencies between fabrics

## Multi-Fabric Hybrid Mode



```
# net file
TCP default
GM 0-63
VAPI 64-127
```

## Multi-Fabric Results



HPL: NP=64; P=Q=8; N=101528; NB=104

Device	$R_{max}$	$R_{peak}$	%
GM	392	266	67.9
TCP	392	236	60.2
Hybrid	392	265	67.6

## Selected Performance Results

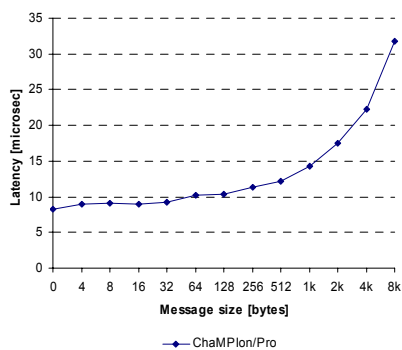
- InfiniBand/VAPI point-to-point and HPL
- Myrinet/GM-2 point-to-point
- Quadrics/ELAN point-to-point
- Point-to-point of Gigabit Ethernet/TCP: ChaMPlon/Pro vs. MPICH
- MPI IO: Mercutio (ChaMPlon/Pro) vs. ROMIO (MPICH)

## LANL InfiniBand Cadillac

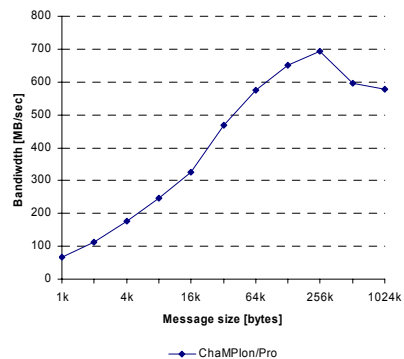
- 128 nodes
- 2.2 GHz Xeon processors
- 3GB RAM
- Linux
- InfiniBand/VAPI
- Used 64 nodes for an HPL run over IB

## Point-to-Point

ChaMPlon/Pro-VAPI Latency



ChaMPlon/Pro-VAPI Bandwidth





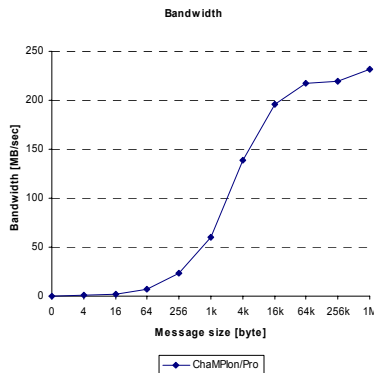
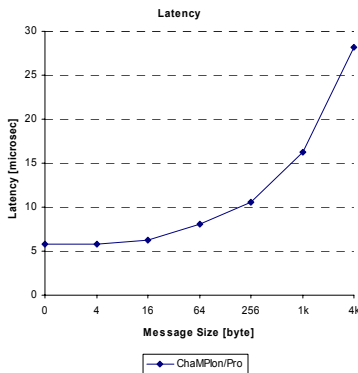
## HPL on LANL Cadillac NP=64

T/V	N	NB	P	Q	Time	Gflops
W00L2L2	120000	104	8	8	5149.31	2.237e+02
Ax-b  _oo / ( eps *   A  _1 * N ) =						0.0217464 ..... PASSED
Ax-b  _oo / ( eps *   A  _1 *   x  _1 ) =						0.0028117 ..... PASSED
Ax-b  _oo / ( eps *   A  _oo *   x  _oo ) =						0.0005249 ..... PASSED

NP = 120000, NB = 104, P = 8, Q = 8  
 $R_{\max}$  = 223.7 Gflops  
 $R_{\text{peak}}$  = 281.6 Gflops  
Efficiency: **80%** of peak

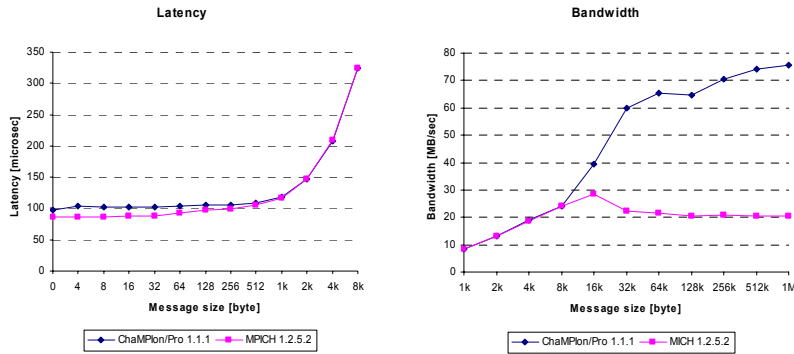


## Point-to-point at LLNL MCR/Quadrics



Xeon 2.40GHz; 4 GB RAM: QsNet I, ELAN 3, PCI 2.1

## Point-to-Point: TCP ChaMPlon/Pro vs. MPICH

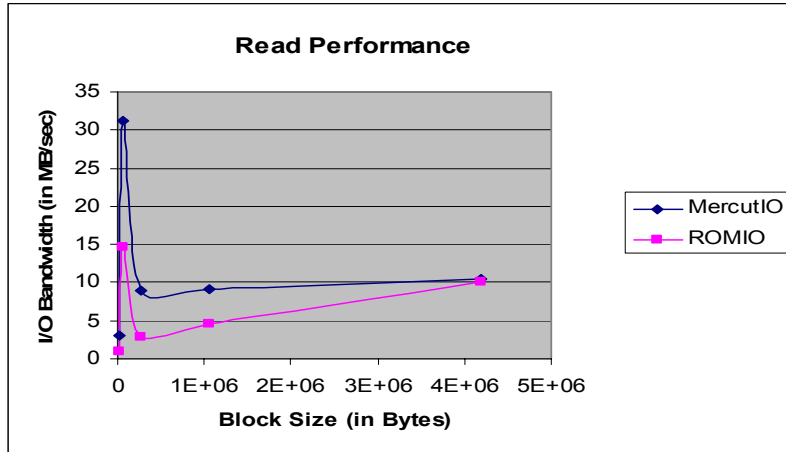


3.06 GHz Xeon; 3 GB RAM; Linux RedHat 9  
Broadcom NetXtreme BCM5704 GigEthernet Controller; bcm5700 driver

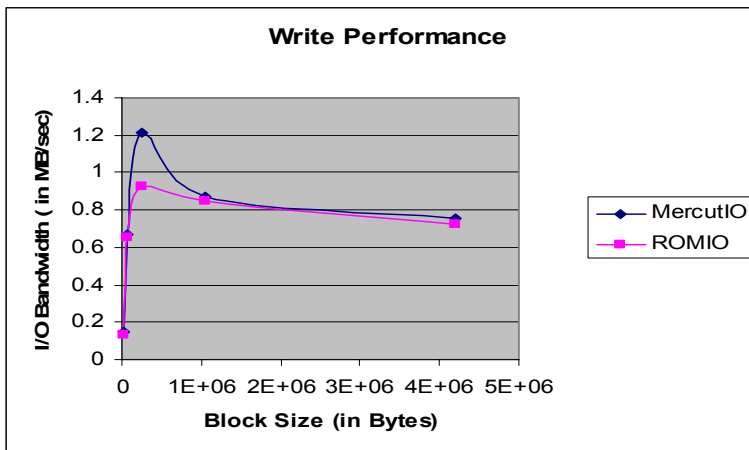
## MercutIO vs. ROMIO

- Hardware Configuration
  - OS: Linux
  - 8 Nodes; 100 Mbps Fast Ethernet
- Software Configuration
  - File system: PVFS 1.5.4
  - ChaMPlon/Pro 1.0 & MPICH 1.2.4
- File access pattern: Contiguous

## MercurtIO vs. ROMIO, II



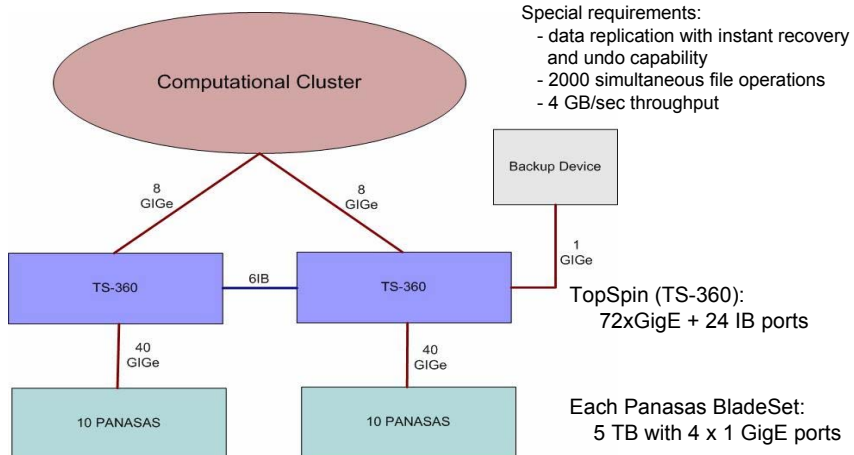
## MercurtIO vs. ROMIO, III



## Terascale Capabilities

- Production installation at NCSA's Myrinet-based Dell 1750 1U 2P Tungsten cluster (Top500 #4)
- 1280 compute nodes + 104 I/O nodes
- 3.06 GHz Xeon processor; 3 GB RAM
- Lustre file system
- 9.73 Tflops on 2,464 processors
- 10.12 Tflops on 2,496 processors
- Proven to work robustly at large scale

## 100 TB Scalable Storage





## Other Verari Cluster Software

- Felix: a comprehensive cluster deployment and management package
- Verari Command Center: cluster management for Verari hardware, including IPMI level monitoring and control
- Cycles@Work: a general purpose job scheduler/resource manager with cycle harvesting capability
- SeeWithin/Pro: MPI 1&2 performance profiler



## Summary

- ChaMPIon/Pro is a full MPI-2 implementation
- Robustness, scalability, performance
- Flexibility, usability, advanced features
- Support for number of target platforms, communication devices, file systems,
- Support for performance monitoring tools, debuggers, and job schedulers
- Emphasis on time to solution
- Ready for end users and integrators/OEMs
- Verari can now deliver a total HPC solution, combining scalable hardware, storage, software, and services



## For More Information

- **Visit**
  - <http://www.verari.com>
  - <http://www.mpi-softtech.com>
- **Contact**
  - [rossen.dimitrov@verari.com](mailto:rossen.dimitrov@verari.com)
- **Visit our publications page**
  - <http://www.mpi-softtech.com/publications>