



CPMD Performance With MPI Collectives Acceleration

March 2011

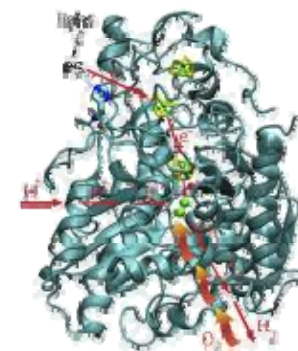
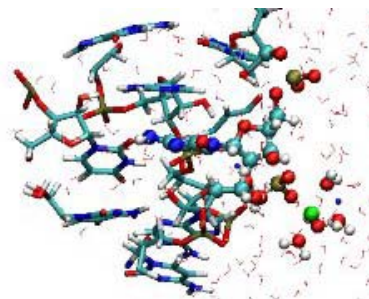
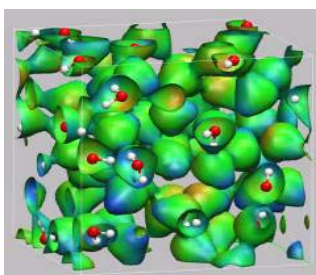


- **The following research was performed under the HPC Advisory Council HPC|works working group activities**
 - Participating vendors: HP, Intel, Mellanox
 - Compute resource - HPC Advisory Council Cluster Center
- **We would like to thank Mellanox for providing early access to its MPI Collectives Acceleration solution (FCA version 2.1)**
- **For more info please refer to**
 - <http://www.hp.com/go/hpc>
 - www.intel.com
 - www.mellanox.com
 - <http://www.cpmd.org>

- **The presented research was done to provide best practices**
 - CPMD interconnect performance benchmarking
 - Application profiling
 - Understanding CPMD communication patterns
- **Preview on available MPI collectives accelerations**
- **First performance results with CPMD**
 - Utilizing MPI collectives accelerations

- **CPMD**

- A parallelized implementation of density functional theory (DFT)
- Particularly designed for ab-initio molecular dynamics
- Brings together methods
 - Classical molecular dynamics
 - Solid state physics
 - Quantum chemistry
- CPMD supports MPI and Mixed MPI/SMP
- CPMD is distributed and developed by the CPMD consortium



Test Cluster Configuration

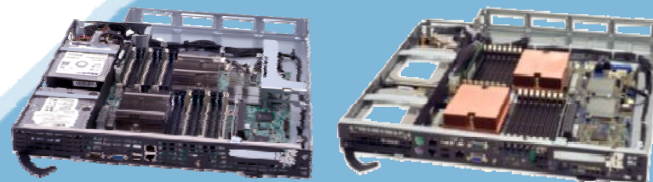
- **HP ProLiant SL2x170z G6 16-node cluster**
 - Six-Core Intel X5670 @ 2.93 GHz CPUs
 - Memory: 24GB per node
 - OS: CentOS5U5, OFED 1.5.2 InfiniBand SW stack
- **Mellanox ConnectX-2 adapters and switches**
- **MPI: Open MPI 1.4.3**
- **Mellanox Fabric Collective Accelerator™ (FCA™) version 2.1**
- **Application: CPMD 3.13.2**
- **Benchmark Workload**
 - C120 - 120 carbon atoms

About HP ProLiant SL6000 Scalable System

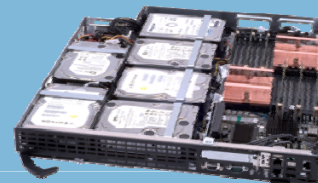
- **Solution-optimized for extreme scale out**



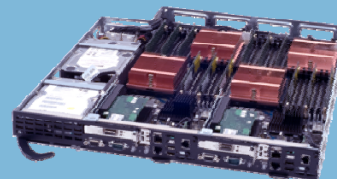
ProLiant z6000 chassis
Shared infrastructure
– fans, chassis, power



ProLiant SL160z G6 ProLiant SL165z G7
Large memory
-memory-cache apps



ProLiant SL170z G6
Large storage
-Web search and database apps



ProLiant SL2x170z G6
Highly dense
- HPC compute and
web front-end apps

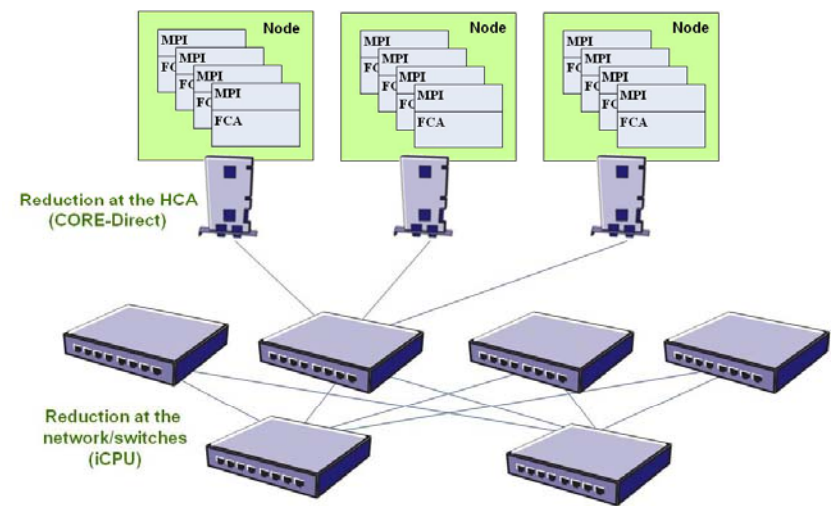
Save on cost and energy -- per node, rack and data center

Mix and match configurations

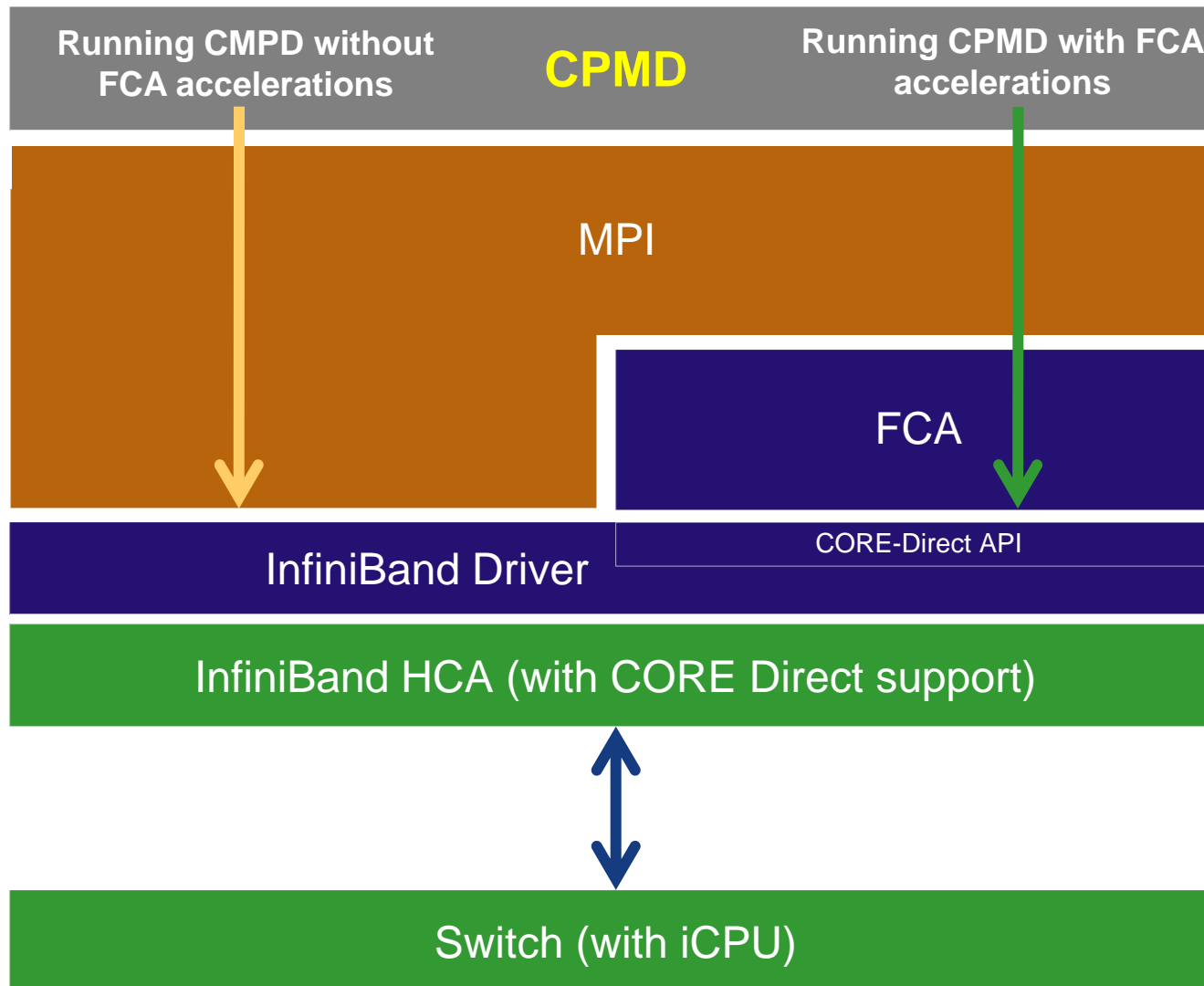
Deploy with confidence

* SPECpower_ssj2008
www.spec.org
17 June 2010, 13:28

- **Mellanox Fabric Collectives Accelerator (FCA)**
 - Utilized hardware accelerations on the adapter (CORE-Direct)
 - Utilized managed switches capabilities (iCPU)
 - Accelerating MPI collectives operations
 - The world first complete solution for MPI collectives offloads
- **FCA 2.1 supports accelerations/offloading for**
 - MPI Barrier
 - MPI Broadcast
 - MPI AllReduce and Reduce
 - MPI AllGather and AllGatherV

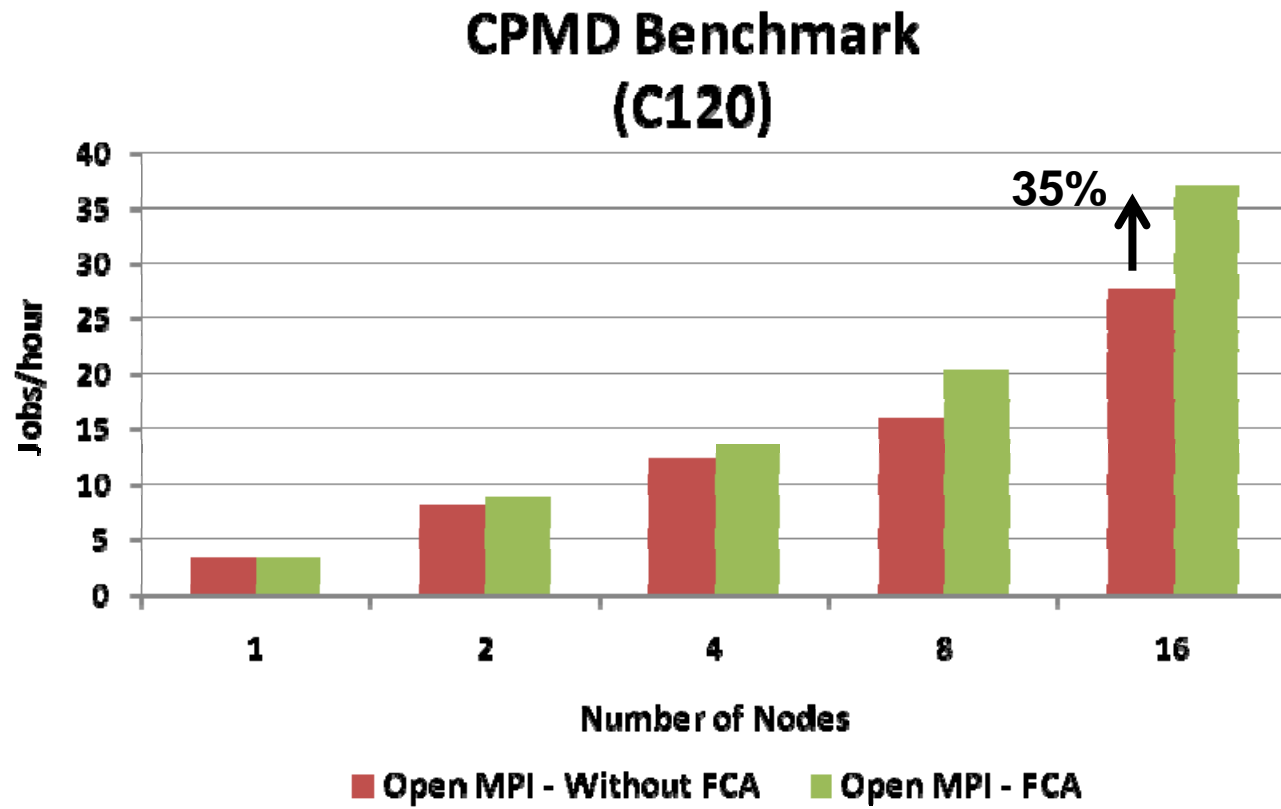


Software Layers Overview



CPMD Benchmark Result – FCA

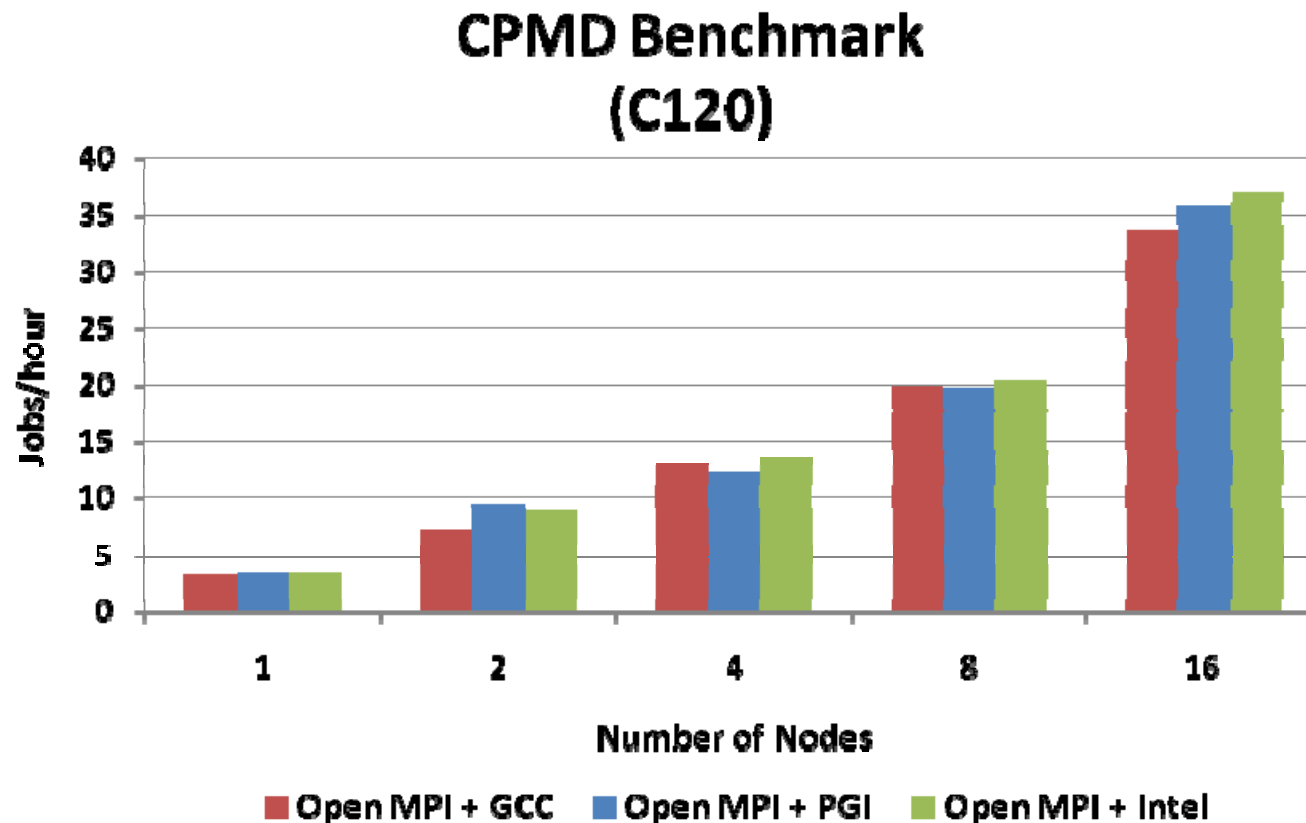
- **FCA accelerates CPMD performance up to 35%**
 - At 16 nodes, 192 cores
 - Performance benefit increases with cluster size – higher benefit expected at larger scale



Higher is better

CPMD Benchmark Result – Compiler

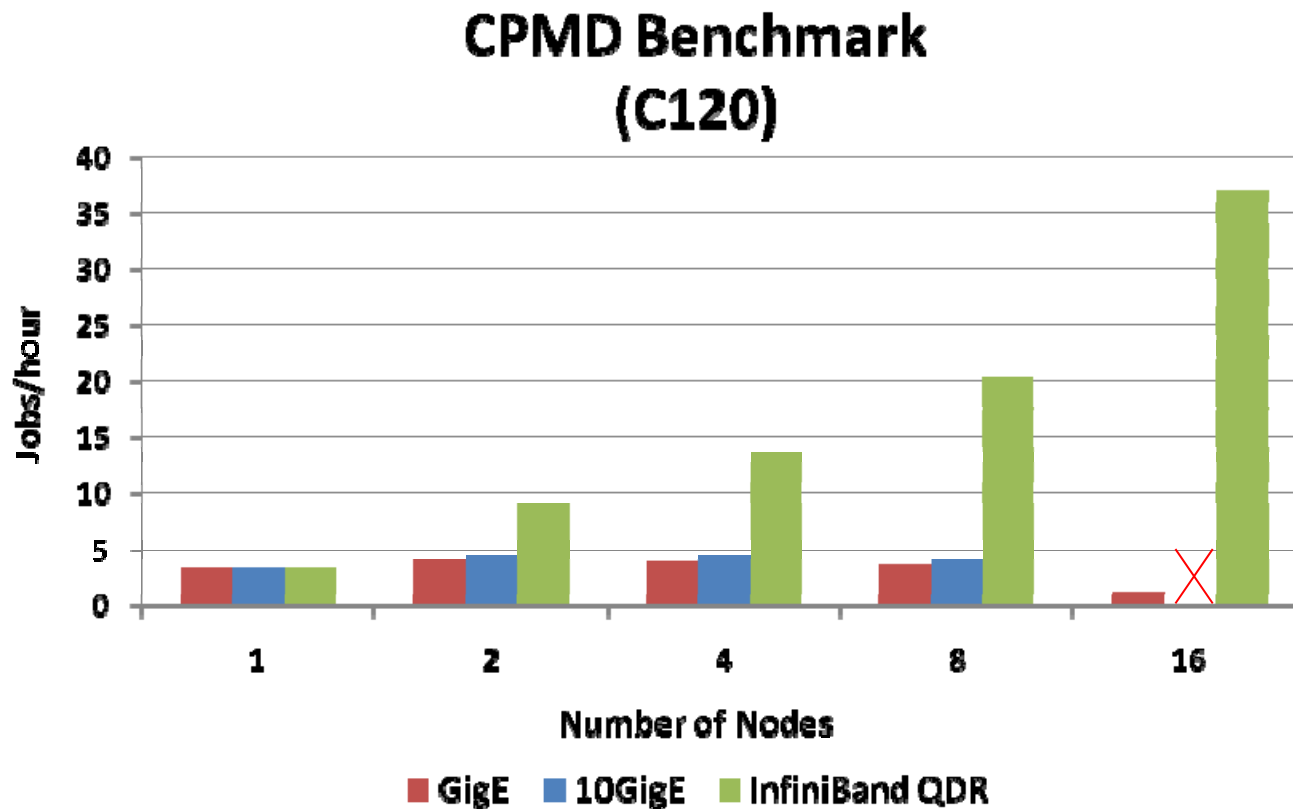
- Intel compiler enables higher performance with CPMD
 - At 16 nodes, 192 cores



Higher is better

CPMD Benchmark Result – Interconnects

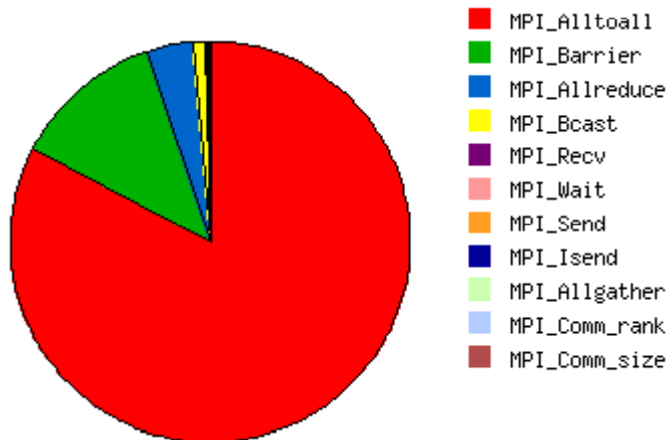
- InfiniBand QDR enables higher application scalability



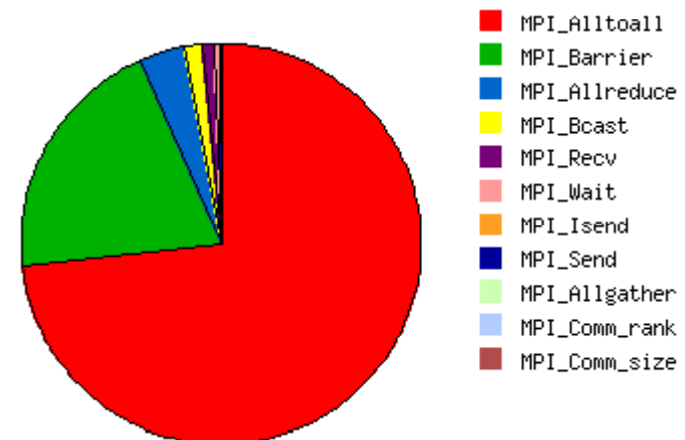
Higher is better

- **MPI collectives generates most communication overhead**
 - MPI_Alltoall, MPI_Barrier, MPI_Allreduce
 - MPI_Barrier overhead increases faster than rest function

96 Ranks

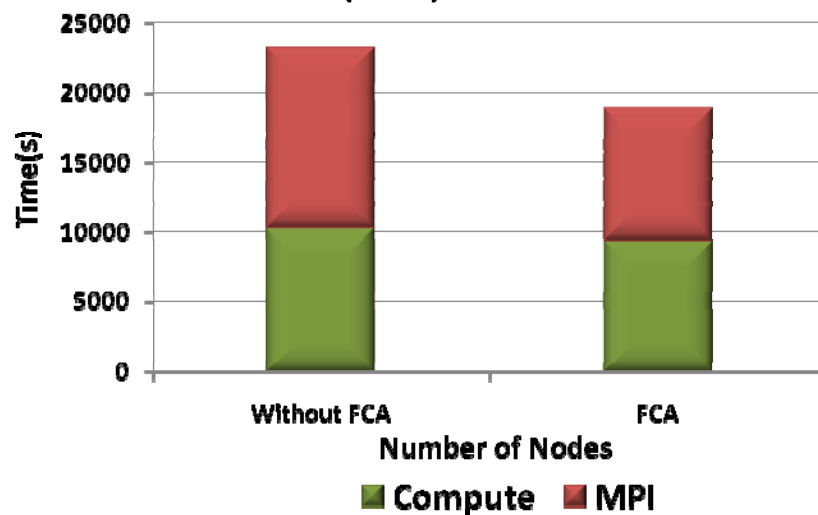


192 Ranks

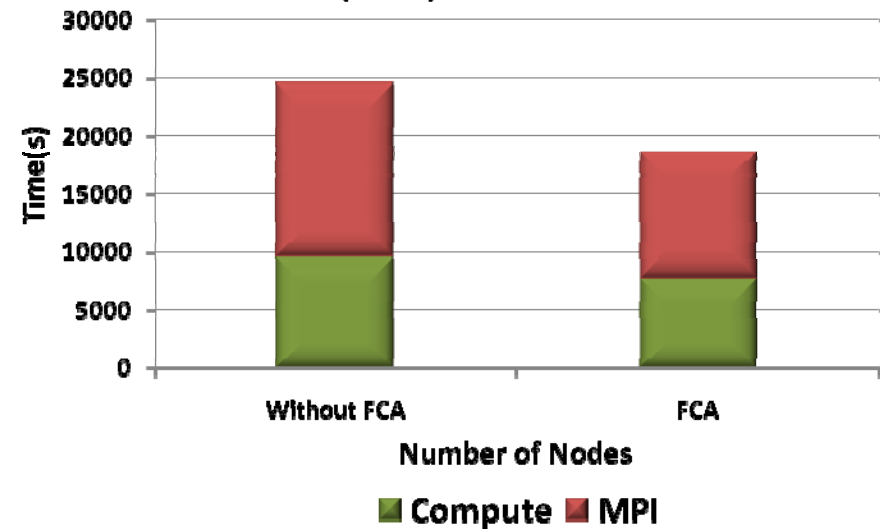


- FCA accelerates application by faster communication/computations

**CPMD Benchmark
(C120)- 96 Processes**



**CPMD Benchmark
(C120) - 192 Processes**



- **MPI Collectives accelerations can dramatically accelerate HPC applications performance**
 - The data presented here reviewed acceleration at small scale
 - Large scale systems will get bigger benefit from such acceleration
- **CPMD MPI profiling**
 - Alltoall, Allreduce, and Barrier are the main MPI routine impacts CPMD performance
- **FCA package has proven to accelerate application even at small scale**
 - Both communication and computation time can be reduced
 - Nearly 34% at 16 nodes for CPMD
 - Higher performance boost expected at larger scale

Thank You

HPC Advisory Council



All trademarks are property of their respective owners. All information is provided "As-Is" without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council Mellanox undertakes no duty and assumes no obligation to update or correct any information presented herein