

Federation Performance on the ORNL p690 cluster

Patrick H. Worley

Oak Ridge National
Laboratory (ORNL)

August 12, 2004

ScicomP 10

Texas Advanced Computing Center
Austin, Texas

Acknowledgements

- Research sponsored by the Atmospheric and Climate Research Division and the Office of Mathematical, Information, and Computational Sciences, Office of Science, U.S. Department of Energy under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.
- These slides have been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes
- Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the United States Department of Energy under Contract No. DE-AC05-00OR22725.

Cheetah



4.5 TFLOPS IBM Power4 Regatta Cluster

27 “Turbo” p690 Nodes

- 32 1.3 GHz Power4 processors per node
- 32 GB on most nodes
 - 5 nodes with 64 GB, 2 with 128 GB

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY


UT-BATTELLE

Cheetah

Original Configuration (12/2001)

- Dual-plane SP Switch2 (Colony) switch
- 19 32-way p690 nodes, each with 2 Corsair network adapters
- 32 8-way LPAR nodes (from partitioning 8 p690s), each with 2 Corsair network adapters, for a total of 8 adapters per p690
- Not using large pages

Current Configuration (2/2004)

- High Performance Switch (Federation)
 - 8 NSB and 4 ISB switches
- 27 32-way p690 nodes, each with 2 2-link network adapters
- 8 4-way p655 nodes (1.7 GHz Power4+ processors), each with 1 2-link network adapter (primarily used for GPFS)
- Using minimum number of large pages per node (only enough to satisfy Federation requirements)

Evaluation Goals

- Determine performance change in moving to Federation from Colony on p690 cluster
 - Communication microbenchmarks
 - Application benchmarks representative of ORNL workload to provide guidance to users on what they should expect.
- Identify any changes in how to most efficiently use the p690 cluster.

Experiment Particulars

- Describing data collected 7/28/2004 - 8/9/2004 using June GA software stack and firmware update (PE Service Pack 7, including PE 4.1.0.7)
- Comparing with results collected in February and March, 2004, showing performance of Federation with earlier software releases, and results collected in 2003 using Colony.
- Default environment variables except for ...
 - export MP_SHARED_MEMORY="yes"
 - export MEMORY_AFFINITY="MCM" (for AIX 5.2)
- Loadleveller script requests
 - #@ network.MPI = sn_all,shared,US
 - #@ node_usage = not_shared

Caveats

- Federation performance still improving, e.g.
 - No striping across adapters yet (still using MAX_PROTO_INSTANCE =1).
- AIX 5.2 performance still changing.
 - Have not yet installed recent efix to improve performance of MP_TASK_AFFINITY. In consequence, running benchmarks both with and without explicit process and thread binding (and not setting MP_TASK_AFFINITY).
- Not using large pages.

Outline

Application Benchmarks

- GYRO fusion code: (bandwidth sensitive)
- POP ocean code (latency sensitive)
- CAM atmospheric code (sensitive to everything)

Latency and Bandwidth

- MPI SWAP benchmarks

GYRO

- GYRO is an Eulerian gyrokinetic-Maxwell solver developed by R.E. Waltz and J. Candy at General Atomics. It is used in the DOE SciDAC Fusion Energy project studying plasma microturbulence.
- Primary communication cost in this benchmark is calls to MPI_ALLTOALL to transpose distributed matrices.
- Compiled/linked with -q32, -bmaxdata:0x80000000 (-q64 performance is slightly worse.)
- MPI only (no OpenMP)

GYRO Experiment Particulars

Two benchmark problems, both time dependent:

1. GTC.n64.500

- 64-mode adiabatic electron case. It is run on multiples of 64 processors. Duration is 3 simulation seconds, representing 100 timesteps.

2. BCY.n16.b.25

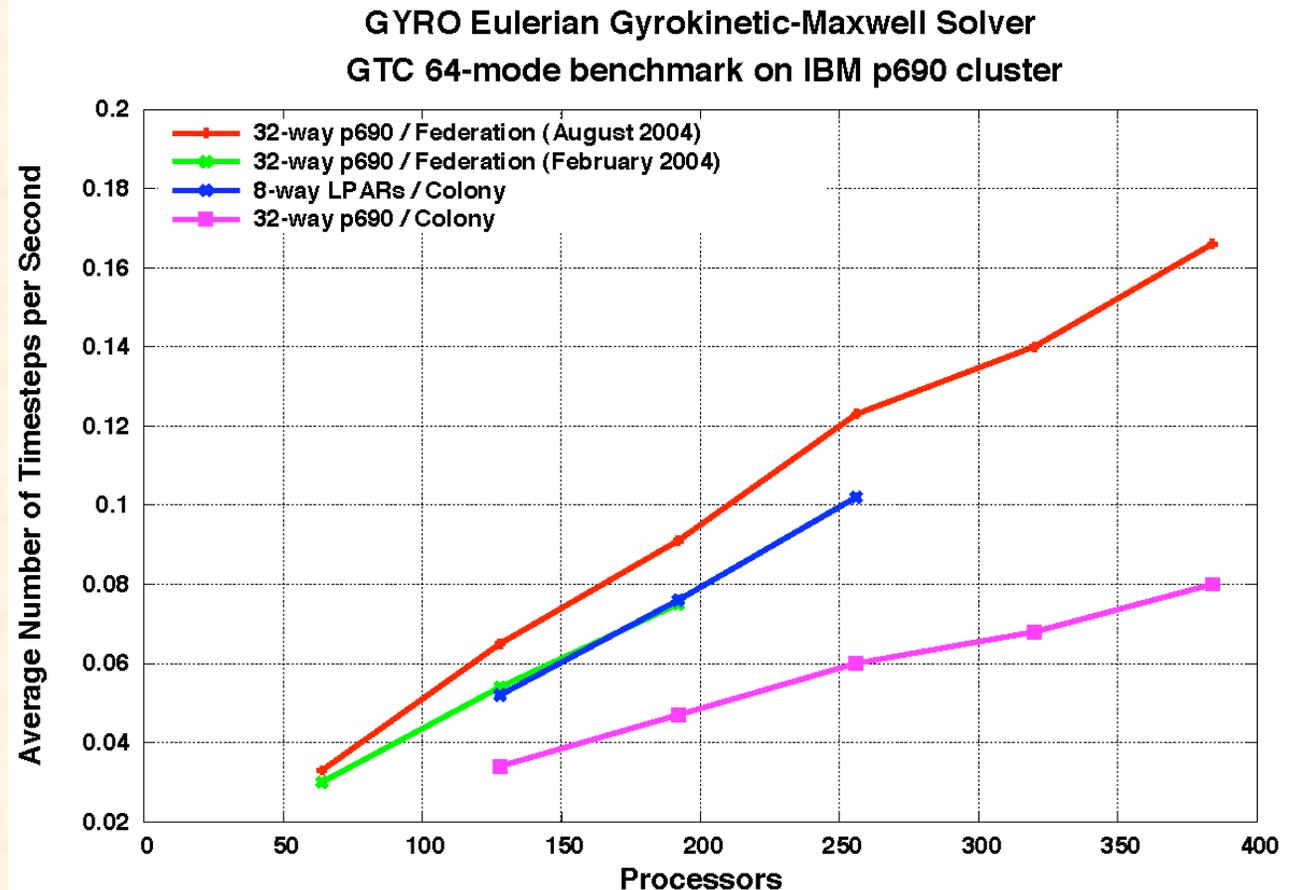
- 16-mode electromagnetic case. It is run on multiples of 16 processors. Duration is 8 simulation seconds, representing 1000 timesteps.

(Current production runs use 32 modes, so benchmark #1 is very large, while benchmark #2 is somewhat small.)

GYRO Simulation Rate: GTC

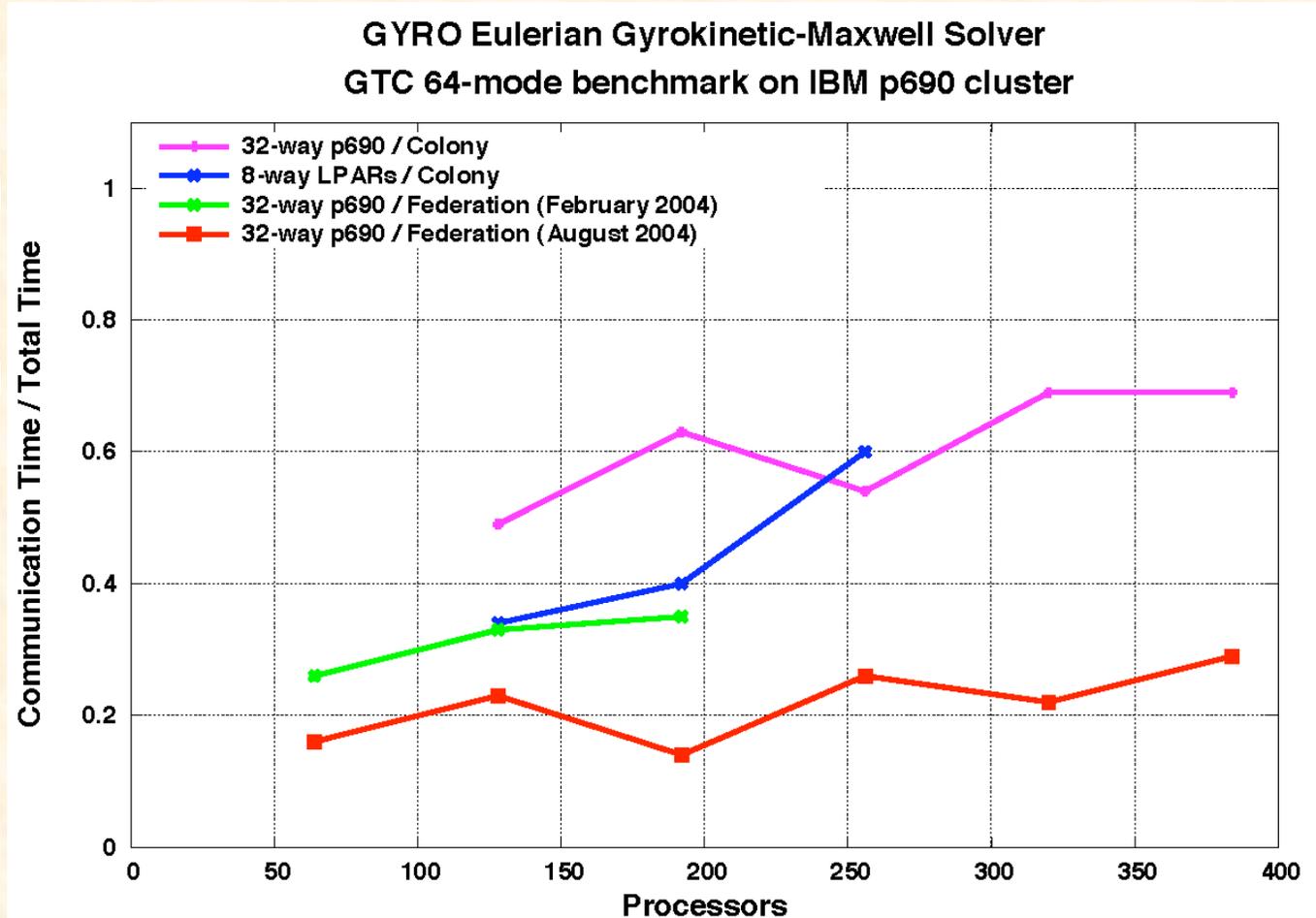
Comparing performance and scaling between Colony and Federation.

- Federation (August 04) is a significant perf. enhancement.
- Before recent firmware upgrade, LPAR/Colony performed as well as Federation.
- Process binding is important for performance for AIX 5.2. Serial performance is better under AIX 5.1 (not using binding).



GYRO Communication Overhead: GTC

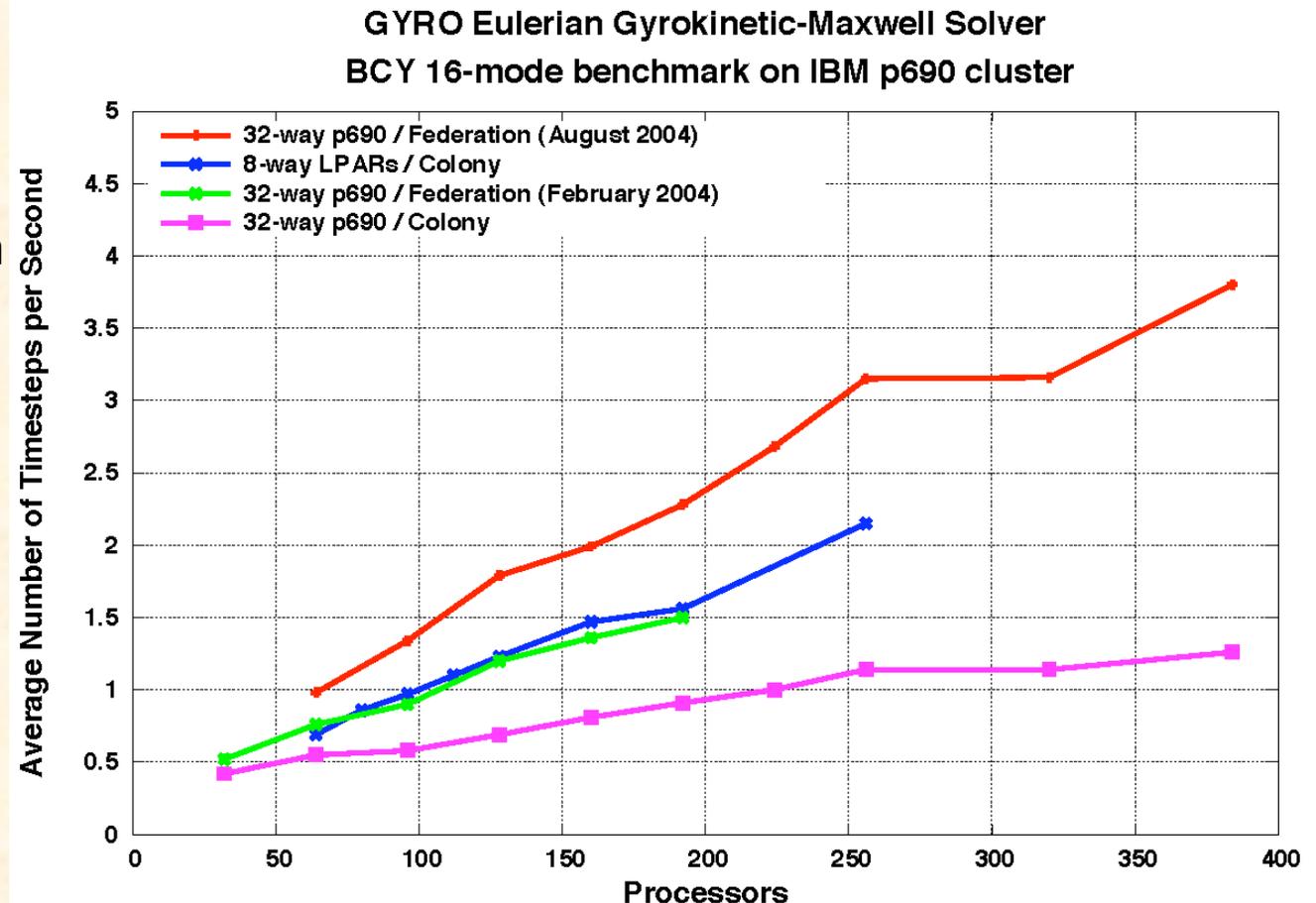
GTC case is bandwidth sensitive. On Colony, communication overhead is as high as 70%. After June update, communication overhead on Federation is never more than 30%.



GYRO Simulation Rate: BCY

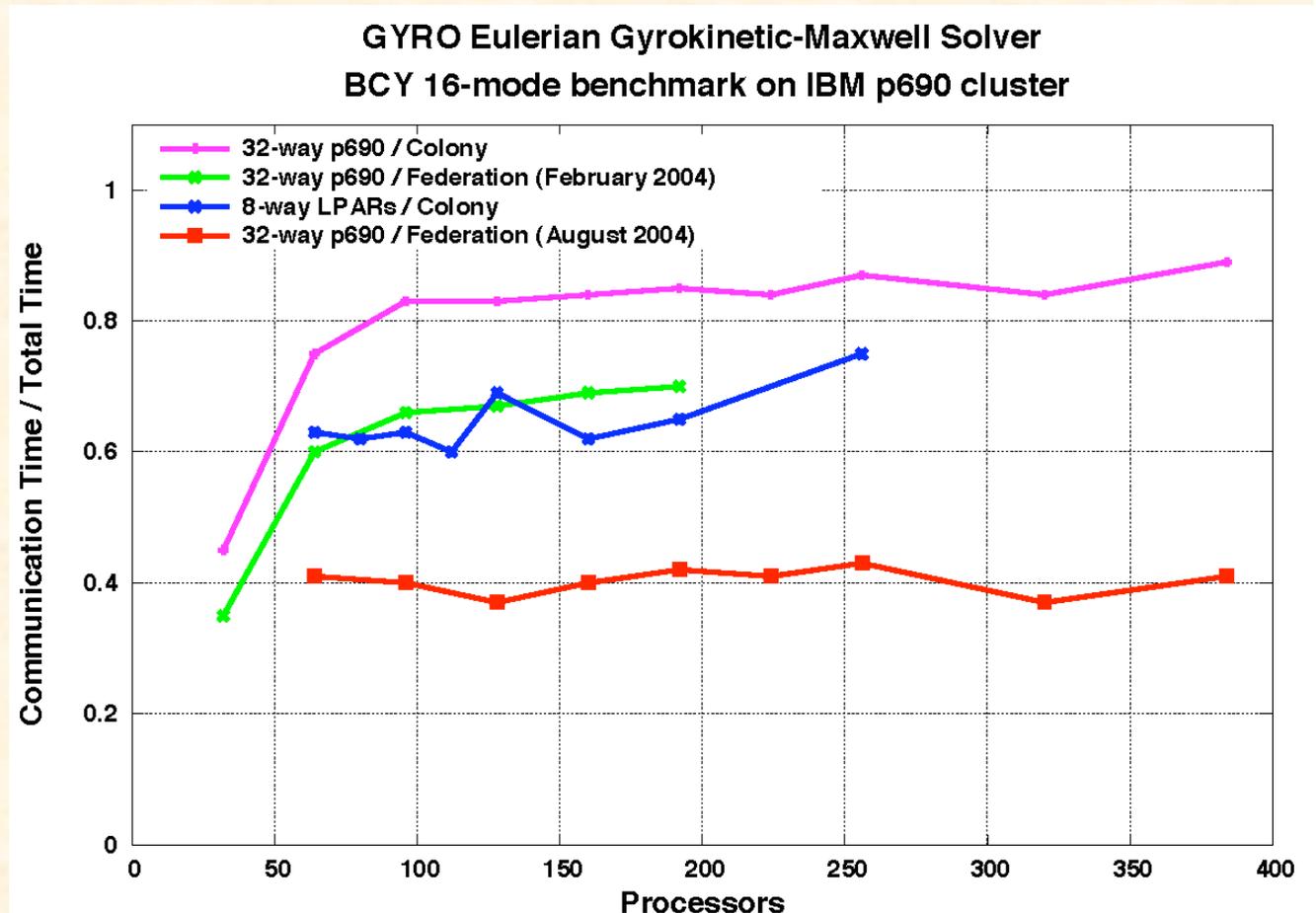
Comparing performance and scaling between Colony and Federation.

- Advantage of Federation (August 2004) is even greater than for GTC.
- Performance is better on AIX 5.2 if not binding processes.



GYRO Communication Overhead

Communication overhead is extremely high for all but Federation (August 2004).



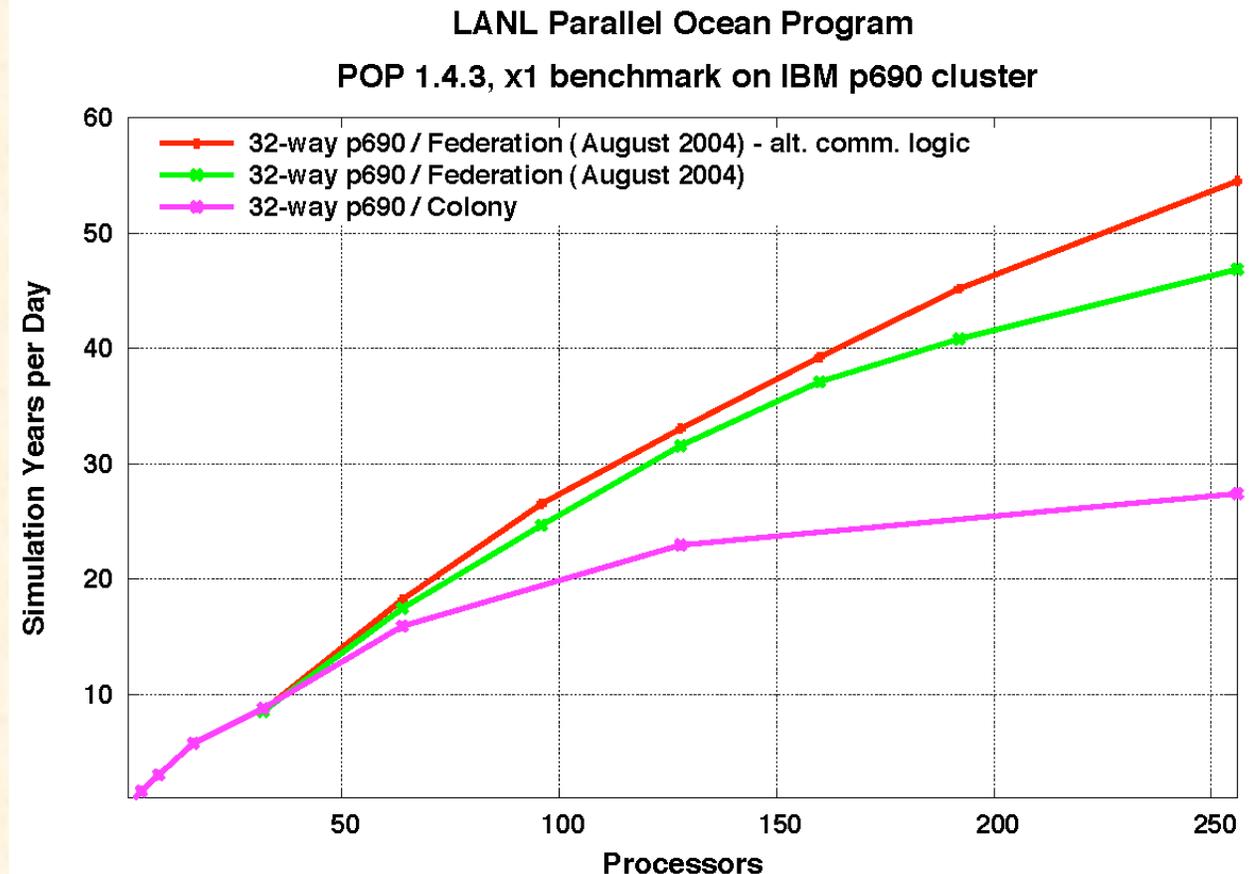
Parallel Ocean Program (POP)

- Developed at Los Alamos National Laboratory. Used for high resolution studies and as the ocean component in the Community Climate System Model (CCSM).
- Two primary computational phases
 - Baroclinic: 3D with limited nearest-neighbor communication; scales well.
 - Barotropic: dominated by solution of 2D implicit system using conjugate gradient solves; scales poorly.
- One fixed size benchmark problem
 - One degree horizontal grid (“by one” or “x1”) of size 320x384x40.
- Domain decomposition determined by grid size and 2D virtual processor grid. Results for a given processor count are the best observed over all applicable processor grids.
- Compiled/linked with -q32, -bmaxdata:0x80000000 .
(-q64 performance is slightly worse.)
- MPI only (no OpenMP)

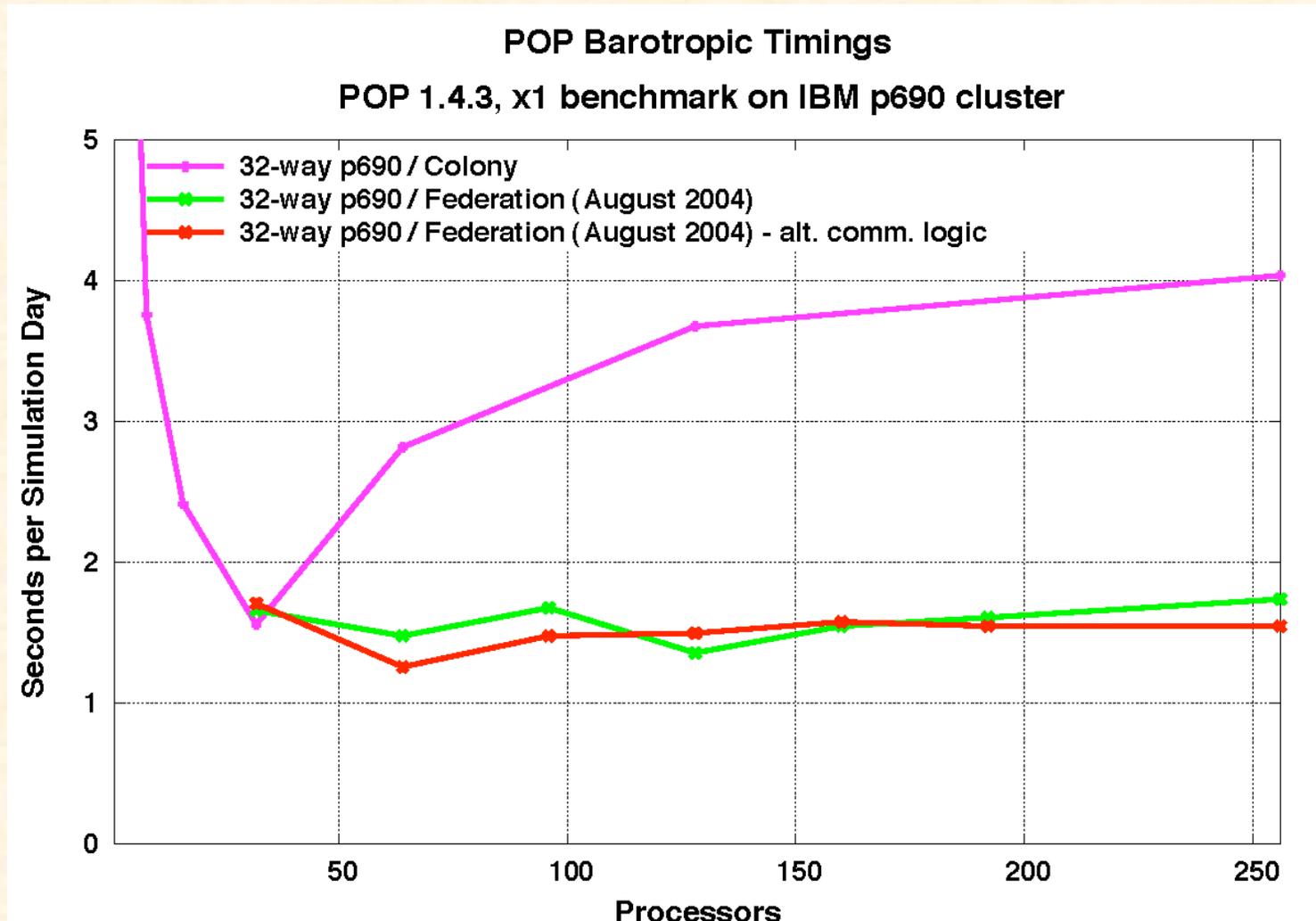
POP Simulation Rate

Comparing performance and scaling between Colony and Federation.

- Alt. Communication logic developed for port to NEC. Works better on most systems, but was not significant on earlier (Colony) experiments.
- Federation nearly doubles the performance of Colony for POP. (POP faster on 32-way p690 / Colony than on 8-way LPAR / Colony.)



POP Performance Diagnosis: Barotropic



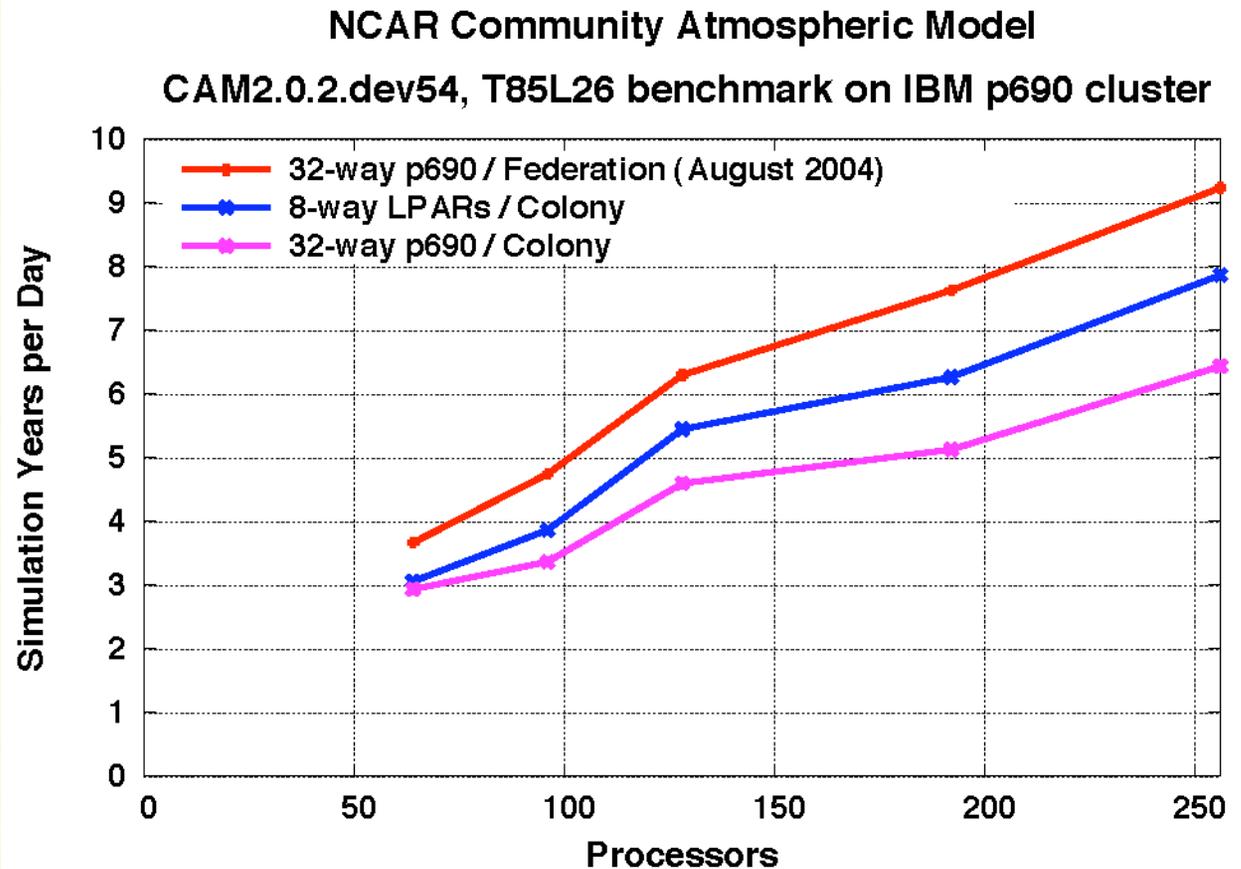
Community Atmospheric Model (CAM)

- Community model whose development is managed by the National Center for Atmospheric Research. Used as the atmospheric component in the Community Climate System Model (CCSM).
- Large code with no simple performance characterization. Suffers from load imbalance that can be addressed statically, but at the cost of additional communication overhead.
- Code has many performance tuning options. Results describe best observed performance over all tuning options.
- Two versions/benchmark problems:
 - CAM2_0_2_dev54, default T85L26 problem, minimal I/O
 - CAM2_0_2_dev68, T85L26 /IPCC problem, full I/O
- Compiled/linked with -q64 .
- Hybrid MPI/OpenMP. Choice of MPI process and OpenMP thread counts for a given processor count is one of the tuning options optimized over.

CAM Simulation Rate: dev54 / no I/O

Comparing performance and scaling between Colony and Federation.

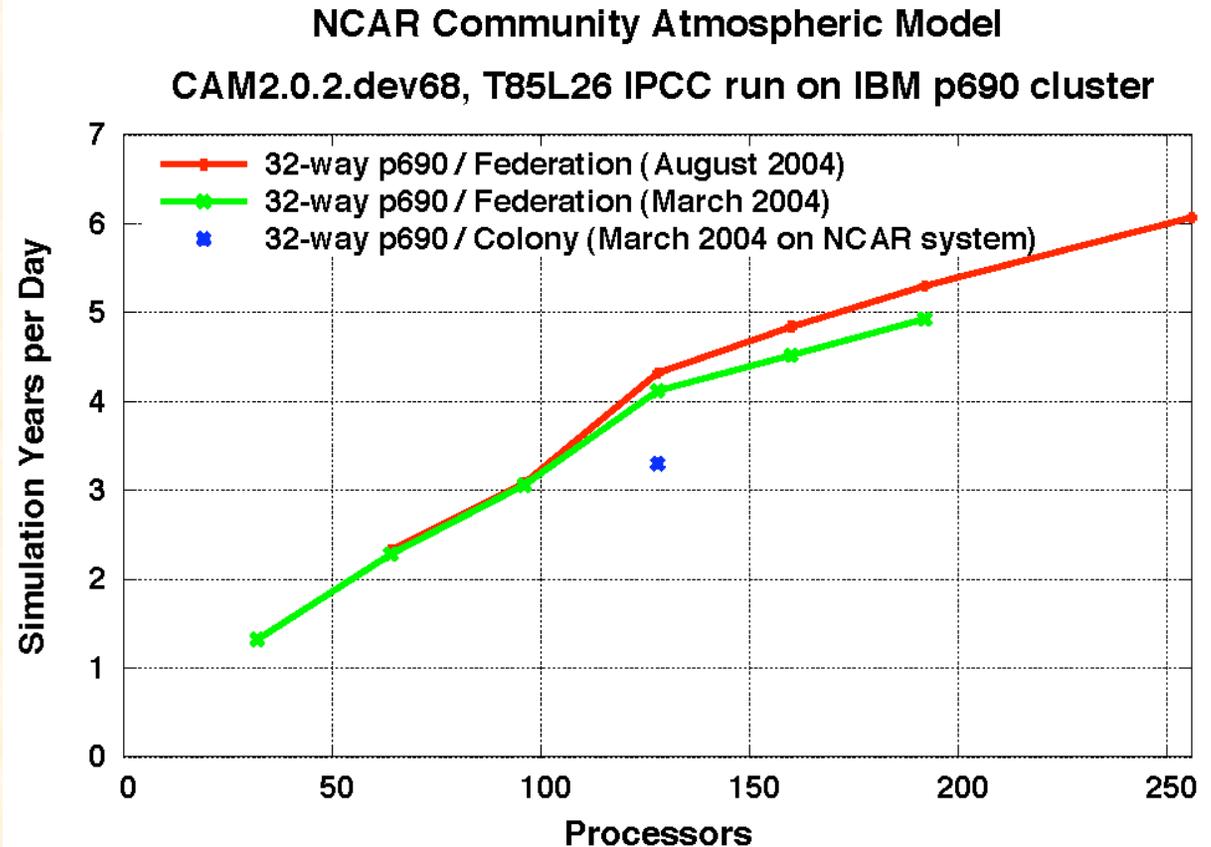
- Federation approx. 40% faster than Colony (32-ways) and 20% faster than Colony (LPARs) for CAM.
- Process/thread binding improves performance.
- MEMORY_AFFINITY=MCM is faster both with and without binding (for 4 OpenMP threads)



CAM Simulation Rate: dev68 / ipcc

Comparing performance and scaling between March and August 2004 on Federation.

- Modest improvement (in real production run) from June update. Good improvement over Colony.
- Different optimal algorithm settings in March and August. In particular, full load balancing (and MPI alltoall) is better than approximate load balancing and point-to-point MPI implementation.



COMMTEST Benchmark

- COMMTEST is a suite of codes that measure the performance of MPI interprocessor communication. In particular, COMMTEST evaluates the impact of communication protocol, packet size, and total message length in a number of “common usage” scenarios. (However, it does not include persistent MPI point-to-point commands among the protocols examined.)
- Compiled/linked with -q64.

COMMTEST Experiments

i-j

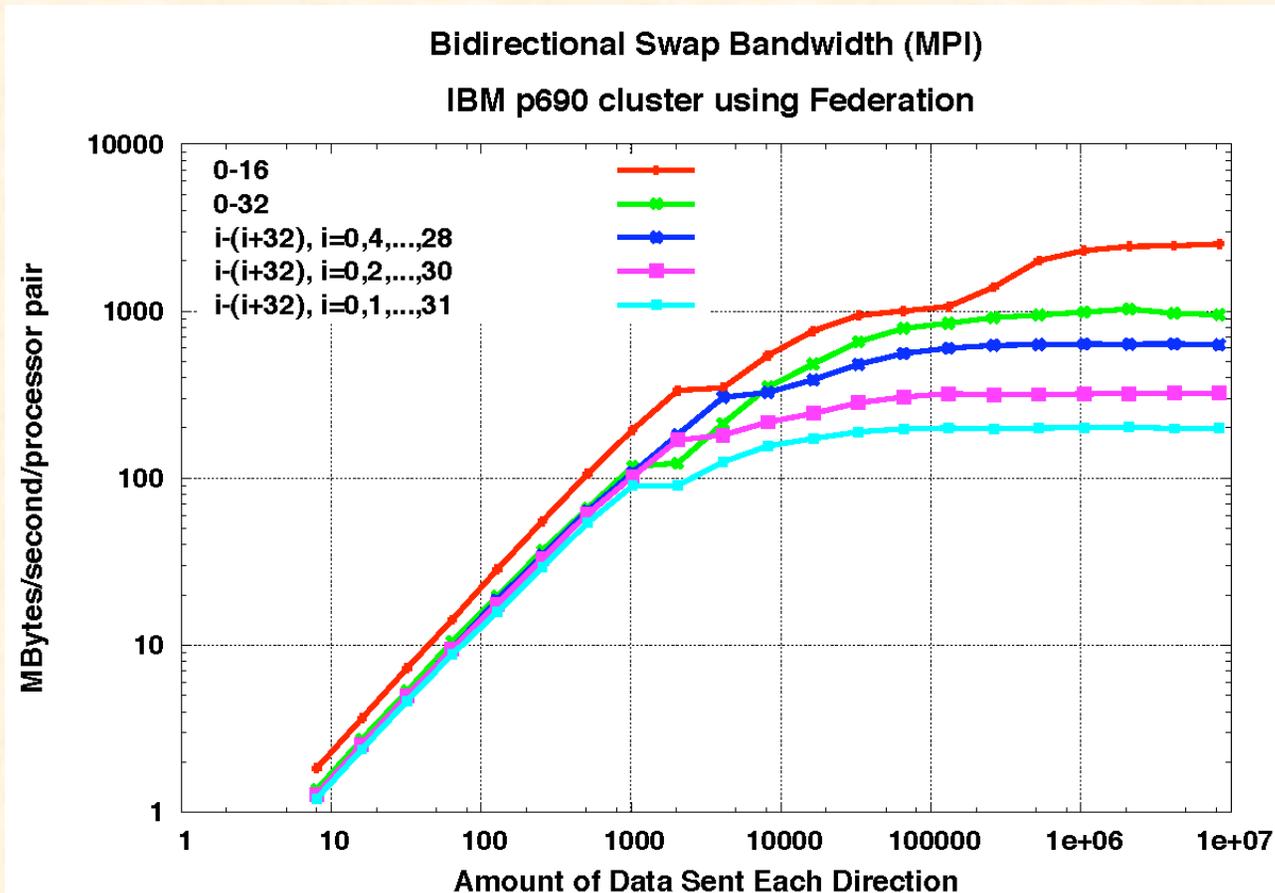
processor i swaps data with processor j. Depending on i and j, this can be within an SMP node or between SMP nodes.

i-(i+j); i=1,...,n; n<j

n processor pairs (i,i+j) swap data simultaneously. Depending on j, this will be within an SMP node or between SMP nodes (or both).

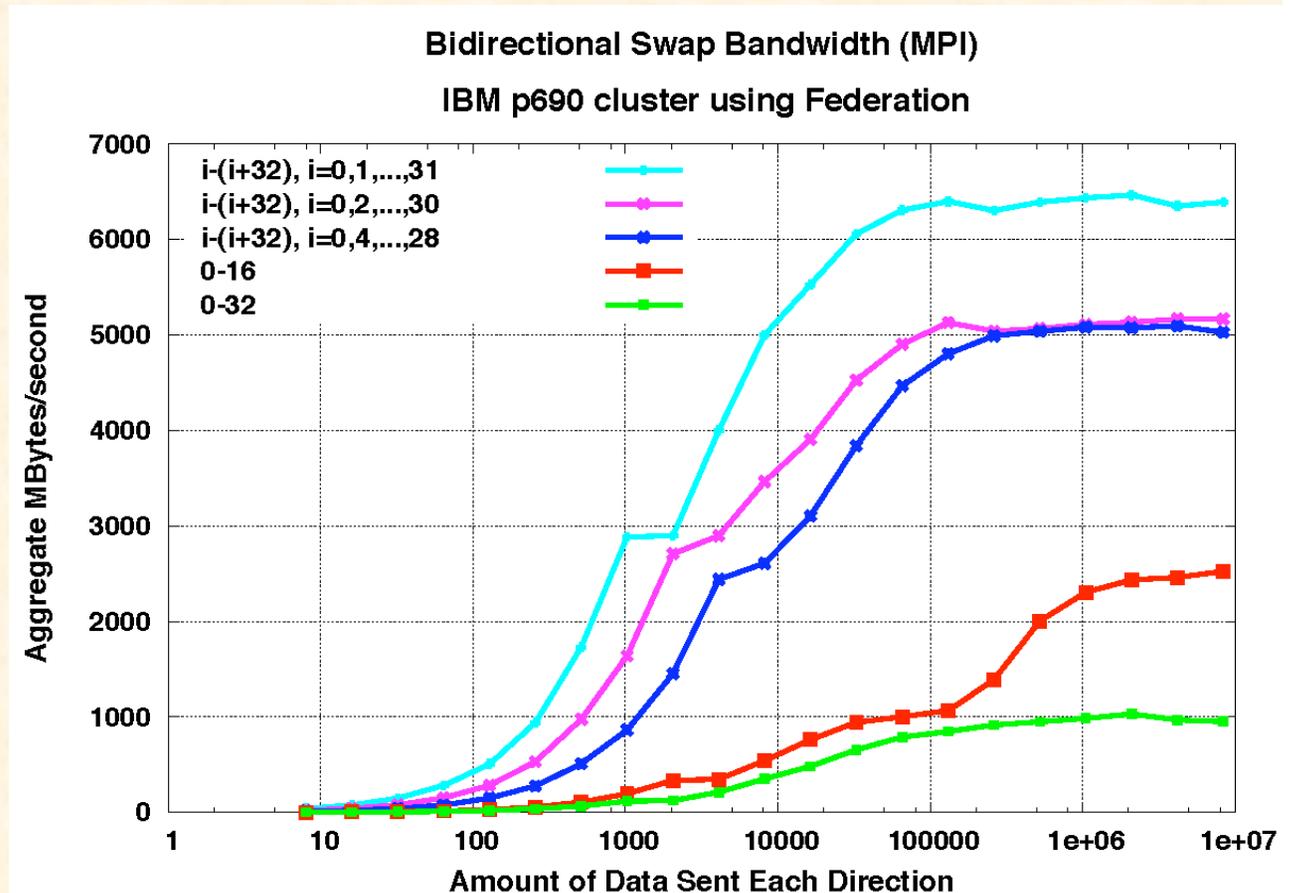
SWAP Benchmark on p690 cluster

Comparing per processor pair performance of SWAP for different communication patterns. Contention for internode bandwidth limits the single pair bandwidth for the simultaneous exchange experiments. All internode experiments have essentially identical small message performance.



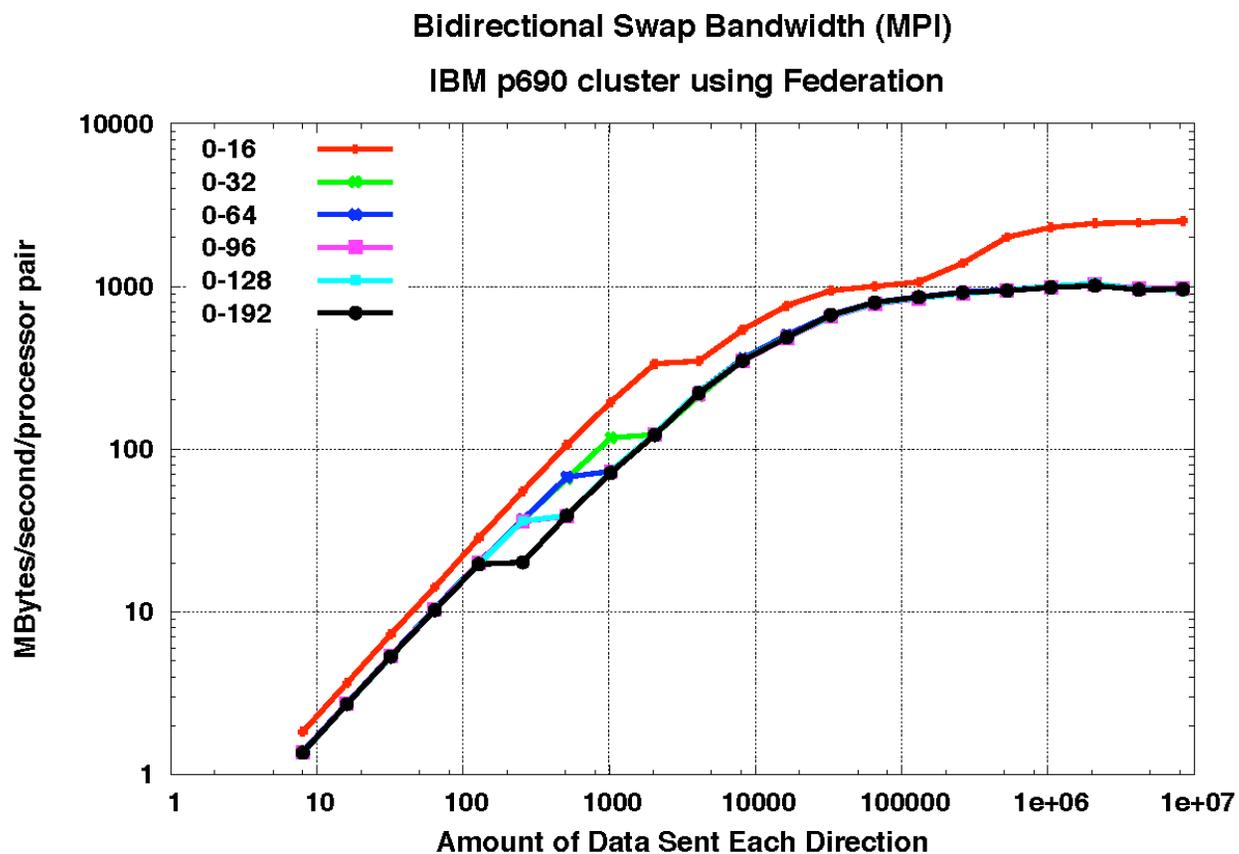
SWAP Benchmark on p690 cluster

Comparing aggregate SWAP bandwidth for different communication patterns. Maximum aggregate bandwidth of 6.5 Gbytes only achieved when all processors communicating.



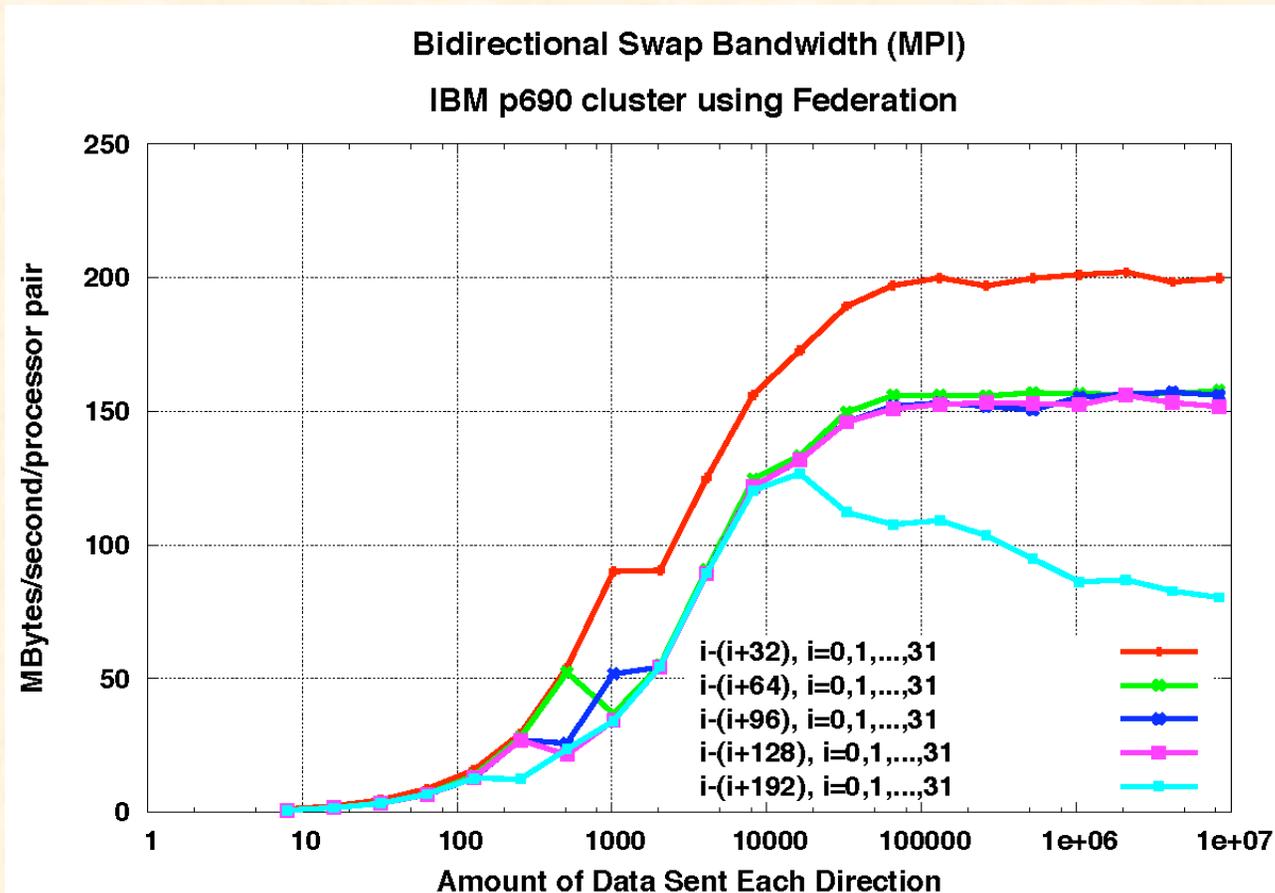
SWAP Benchmark on p690 cluster

Comparing performance of SWAP for a single processor pair for a variety of different total processor counts. The nodes communicating also varies. Performance is identical except for message sizes between 128 and 1024 bytes.



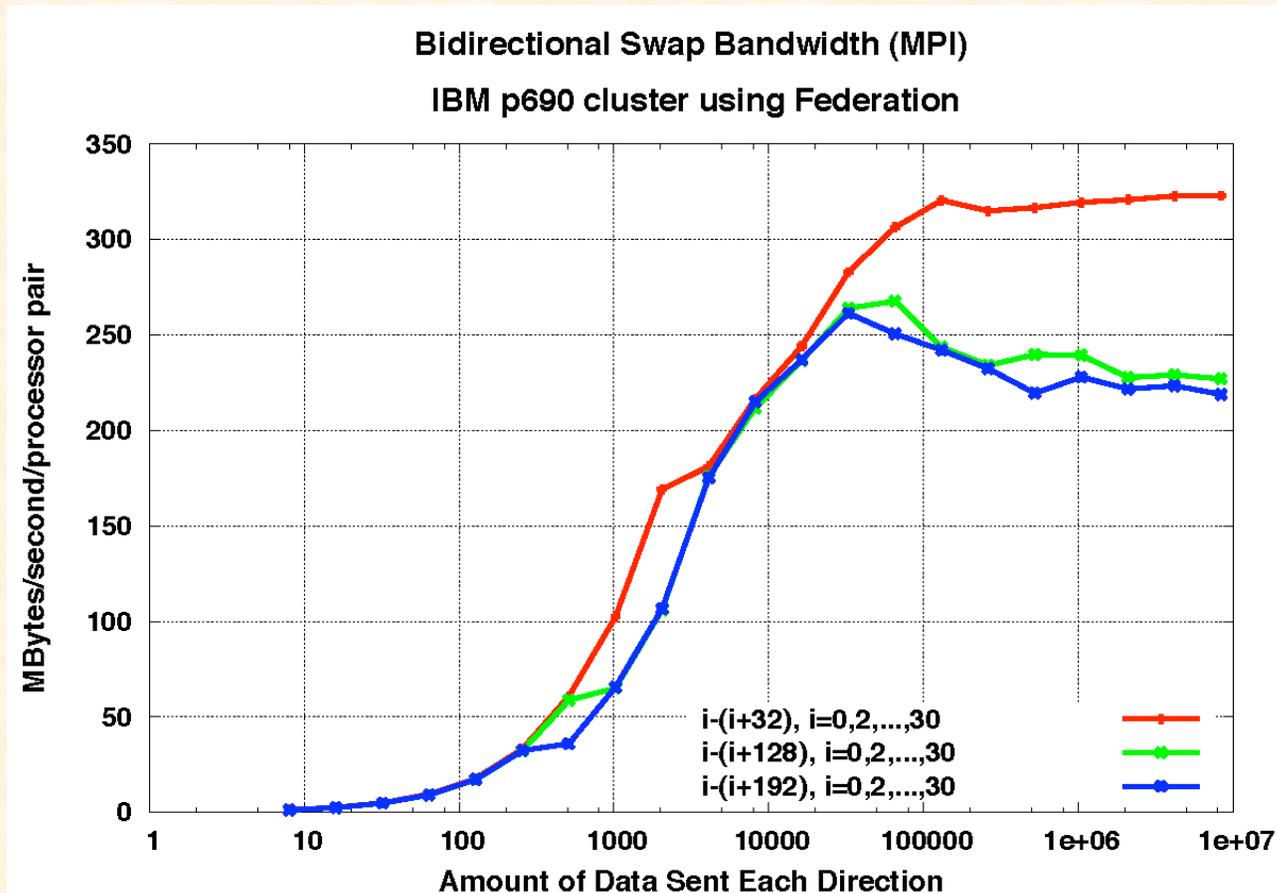
SWAP Benchmark on p690 cluster

Comparing performance of SWAP for 32 simultaneous exchanges between pairs of SMPs for a variety of different SMP node counts. Performance is similar for 4, 6, and 8 nodes. Performance for 2 nodes is significantly better, and performance for 12 nodes is significantly worse.



SWAP Benchmark on p690 cluster

Comparing performance of SWAP for 16 simultaneous exchanges between pairs of SMPs for a variety of different SMP node counts. Performance for 8 and 12 nodes is similar. Performance for 2 nodes is better.



Summary

- Federation has had a significant impact on application performance on the p690 cluster at ORNL.
- Performance characteristics have also changed, requiring re-examination of parallel algorithms.
- There are (still) lots of tuning parameters, including new issues involving MEMORY_AFFINITY, MP_TASK_AFFINITY (or explicit binding), q32 vs. q64, and large pages.
- We expect to see improved performance over next few months with additional updates.
- There are outstanding questions on system scalability. Further experiments are needed to verify and quantify the issues.

Questions ? Comments ?

For further information on these and other evaluation studies, visit

<http://www.csm.ornl.gov/evaluation>

or send e-mail to

worleyph@ornl.gov