Running IFS on an LNXI Opteron cluster at ECMWF

George.Mozdzynski@ecmwf.int

Acknowledgements: Petra Kogel
Sami Saarinen
Peter Towers

Outline

- Motivation
- Opteron and P690+ clusters
- MPI communications
- IFS Forecast Model
- IFS 4D-Var
- Compilers
- Other S/W
- Conclusions

Motivation

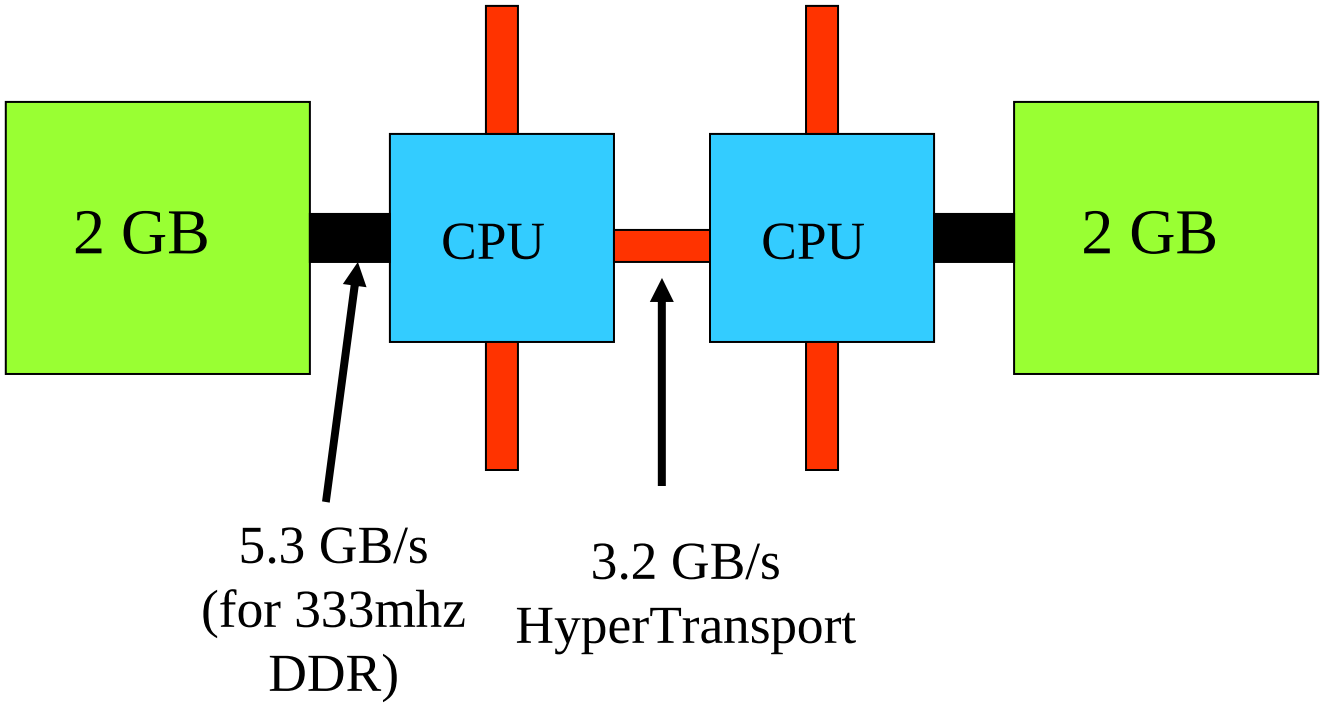
- Linux clusters becoming more popular
- T0P500
  - Half of top 20 systems are Linux Clusters
  - Mostly in research organizations
- NWP centres still using proprietary systems
  - Strict operational schedules
  - Mix of operational and research workloads
- Can Linux clusters be deployed in NWP centres?
- Linux cluster at ECMWF
  - Single node server Oct 2003 (2 CPU Opteron 2.0 GHz)
  - LNXI Cluster Installed 2Q2004

System characteristics

	IBM P690+	LNXI Opteron
Clusters	2 (Dec04)	1
CPUs / cluster	2176	66
Clock	1.9 GHz	2.2 GHz
Peak Gflops / CPU	7.6	4.4
Memory GB / node	32	4
Useable memory / node	24	3.5


Linux Networx
Linux Networx

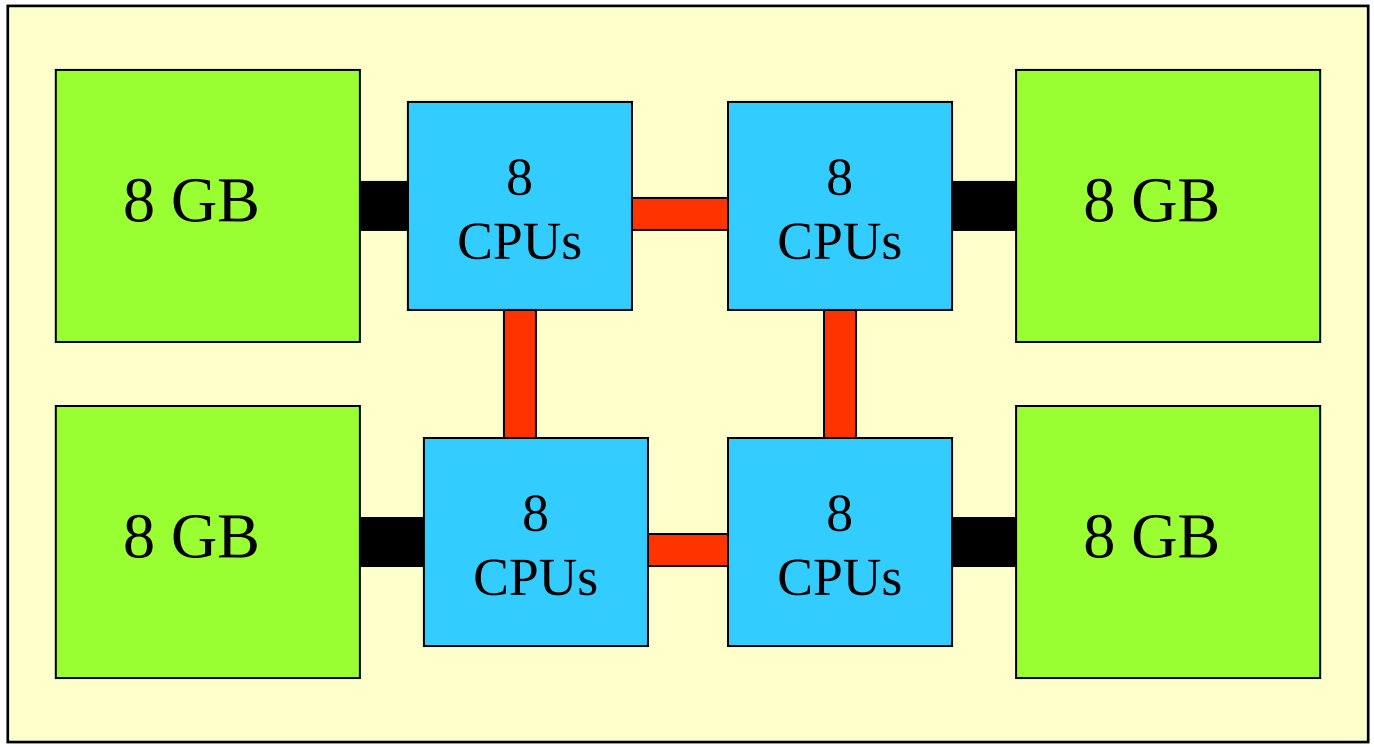
AMD Opteron Architecture (2 CPU node)



IBM P690+ Node

Also local/remote memory

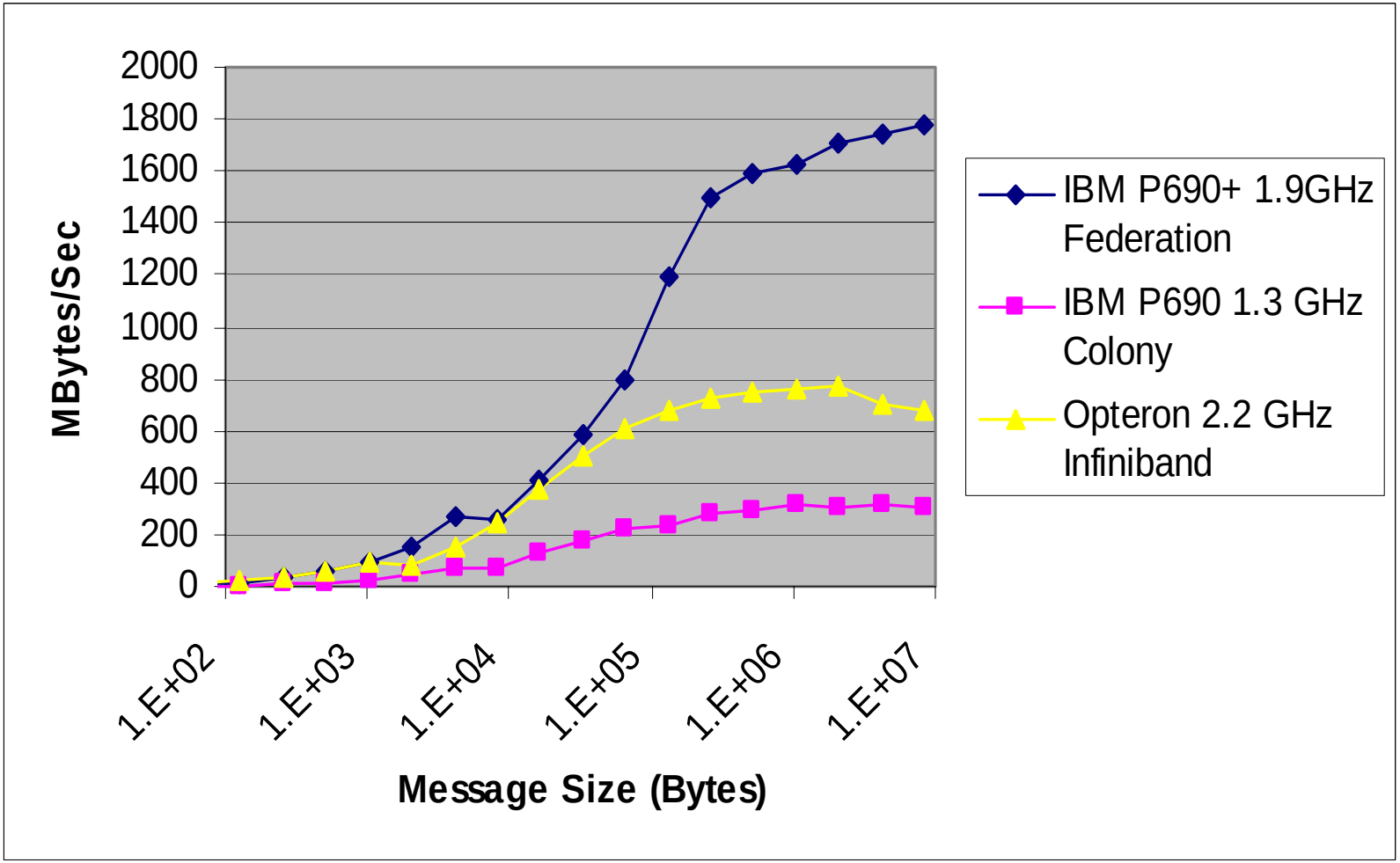
AIX SUPPORTS environment vars
MEMORY_AFFINITY=MCM
MP_TASK_AFFINITY=MCM



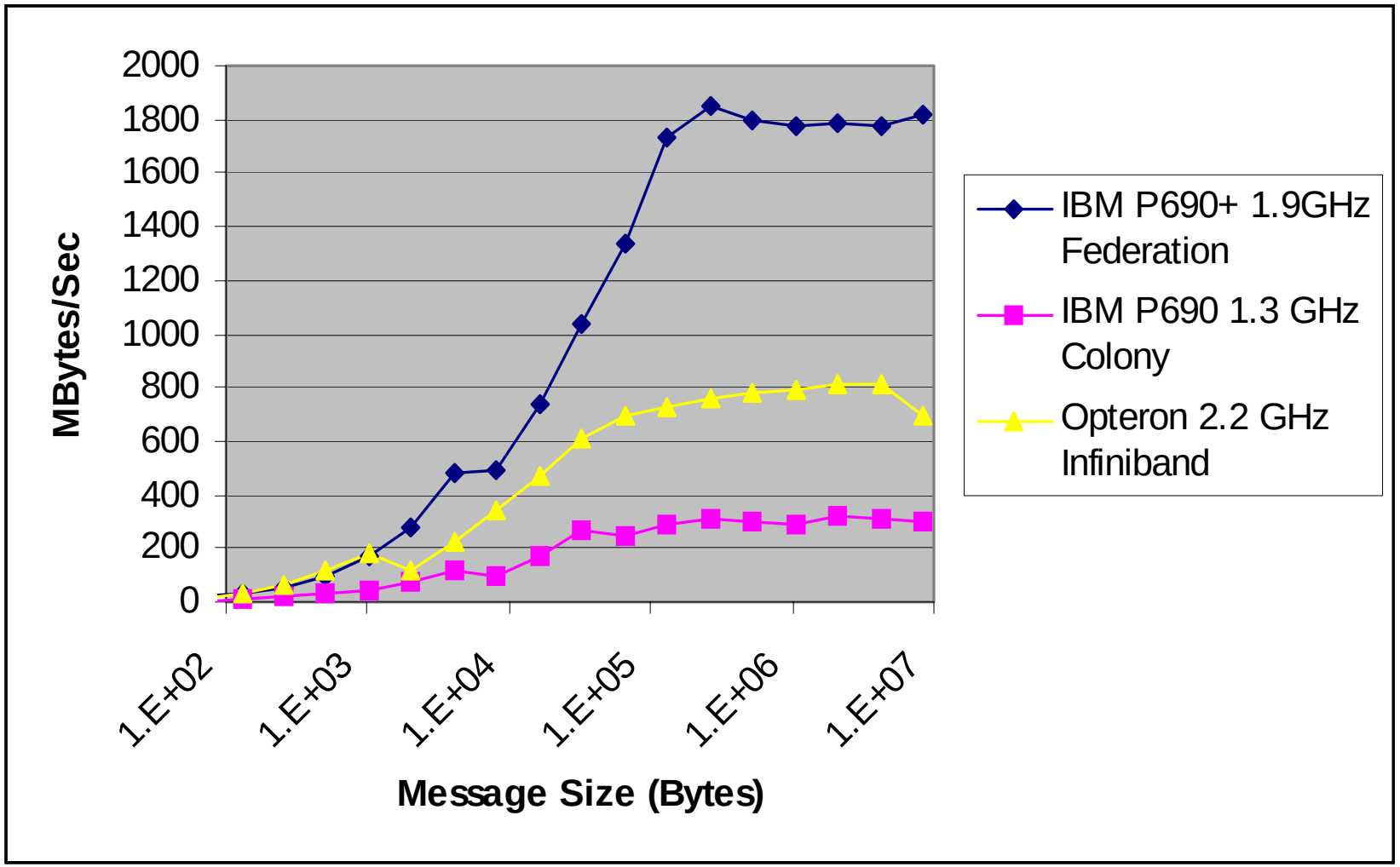
MPI Latency

	Latency micro-secs
Opteron 2.2 GHz Infiniband	5
IBM P690+ 1.9 GHz Federation	6
IBM P690 1.3 GHz Colony	24

MPI Bandwidth (PING-PONG)



MPI Bandwidth (PING-PING)



MPI Bandwidth/CPU

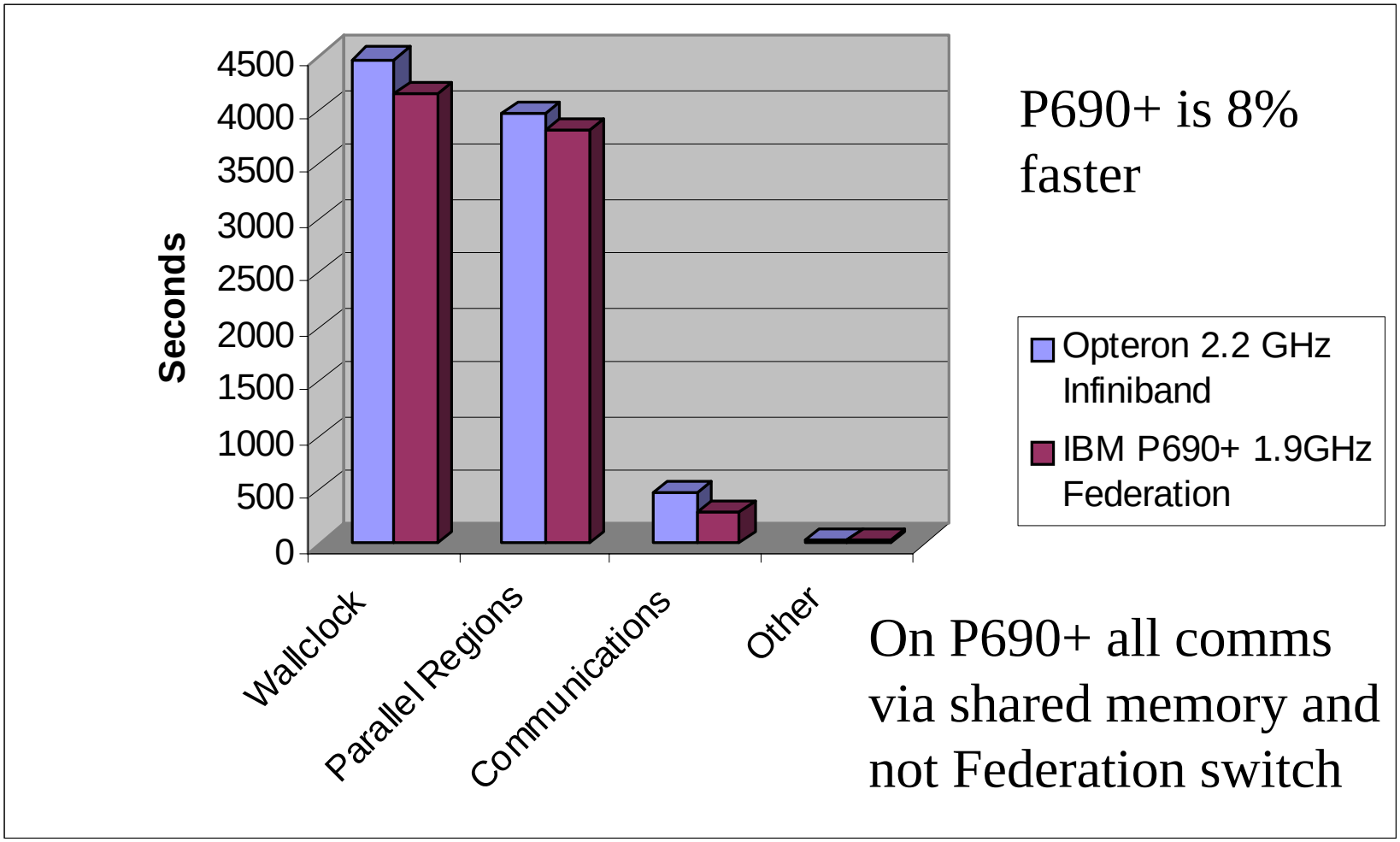
	MB/s /Link	Links /Node	MB/s /Node	CPUs /Node	MB/s /CPU
Opteron 2.2 GHz Infiniband	800	1	800	2	400
IBM P690+ 1.9 GHz Federation	1700	4	6800	32	213
IBM P690 1.3 GHz Colony	320	1**	320	8	40

** one link per 8 CPU LPAR

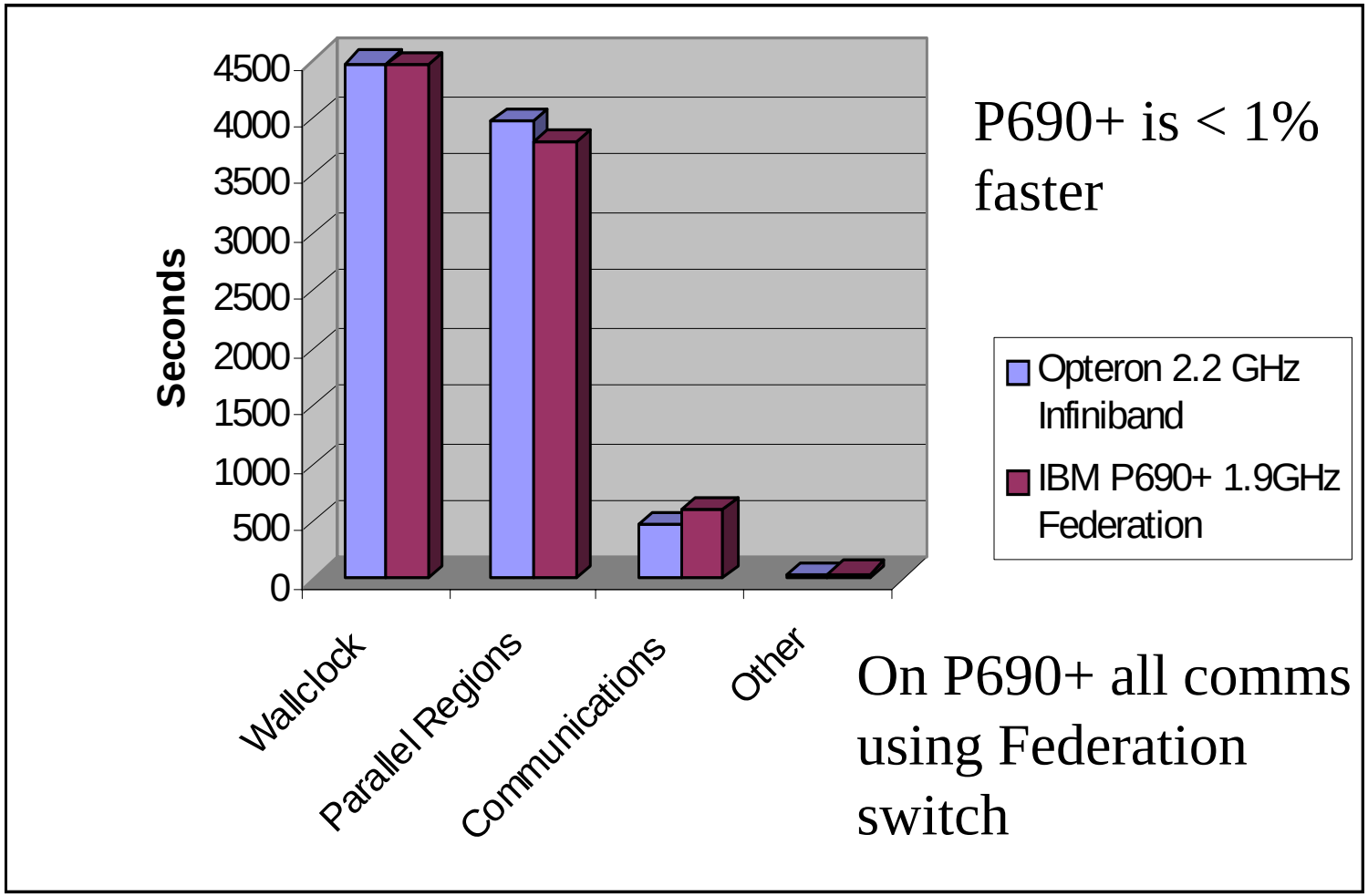
MPI for Infiniband

- **OSU MVAPICH**
  - Used successfully for IFS applications
  - Environment variables not inherited (easily resolved)
- **NCSA MPICH-VMI**
  - Encountered some problems with IFS
- **Open MPI**
  - Preliminary release at Supercomputing 2004
  - Full MPI-2
  - Combines technologies from
    - MVAPICH
    - LAM/MPI
    - Other implementations

RAPS8 IFS Forecast Model T_L399L62 using 32 CPUs



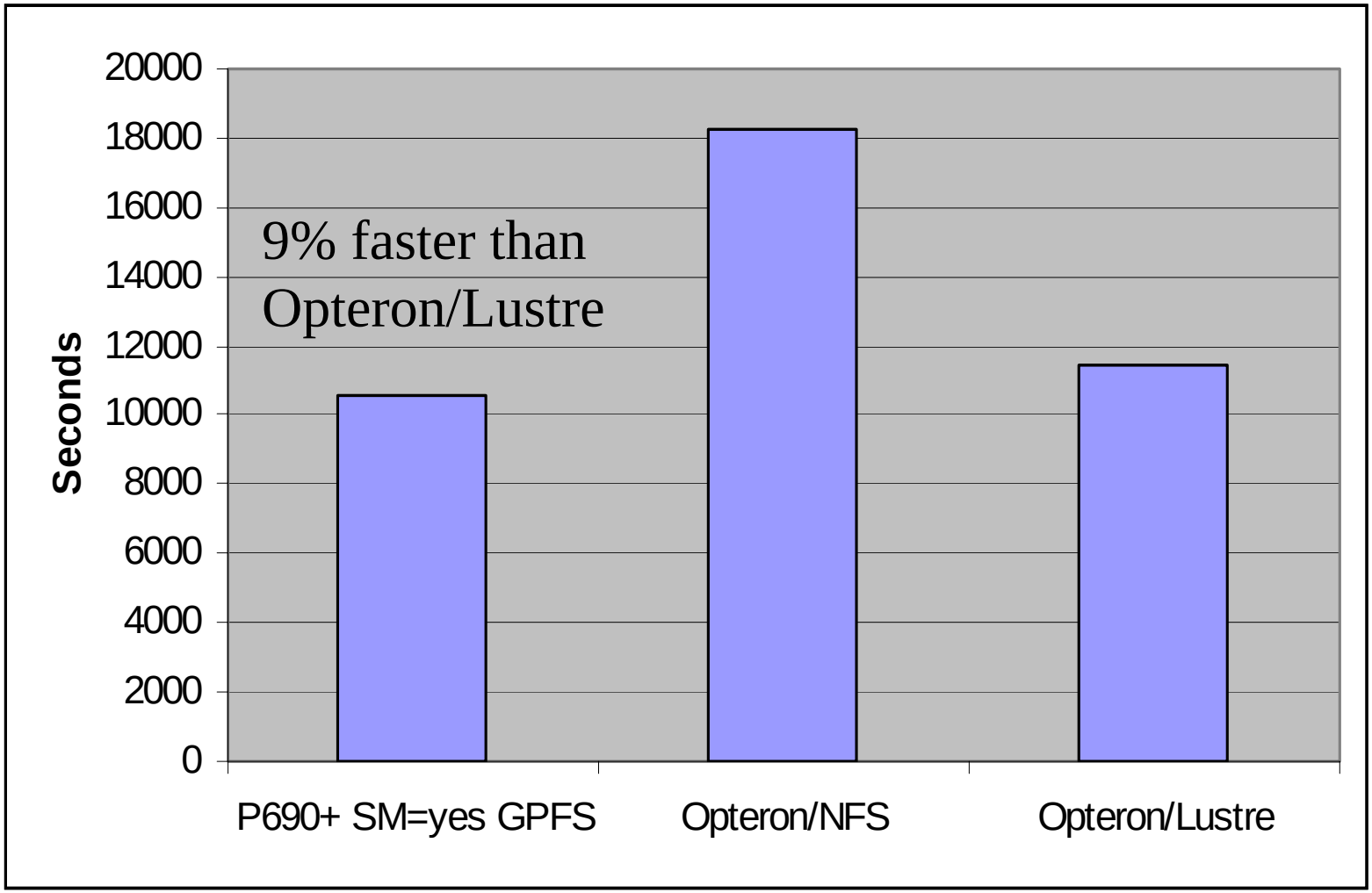
RAPS8 IFS Forecast Model T_L399L62 using 32 CPUs (and Federation switch)



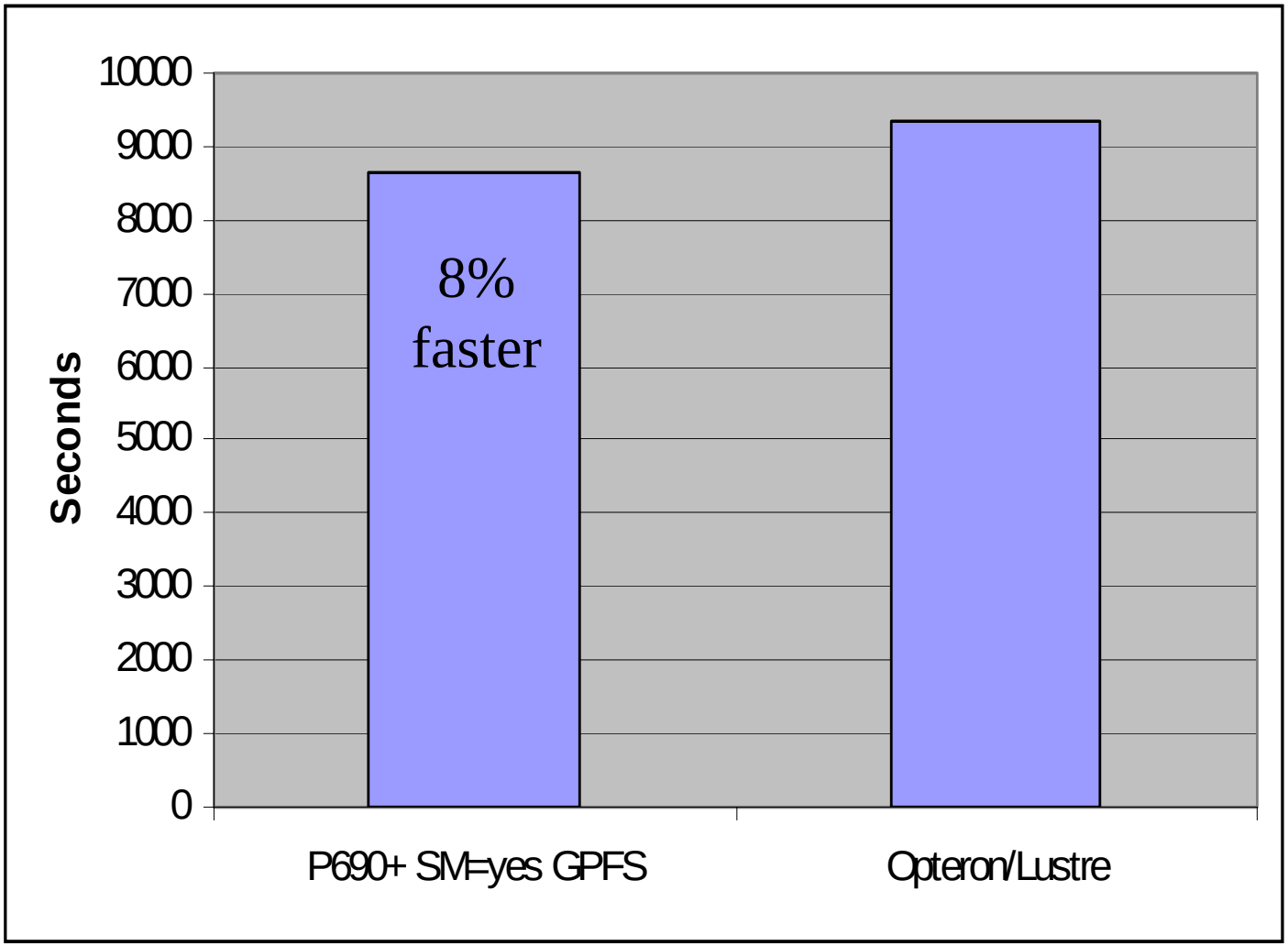
IFS T_L399L62 Model

- IBM P690+ 1.9 GHz CPU (4 flops per clock)
  - 7.6 Gflops peak, 670 Mflop sustained (9% of peak)
- AMD Opteron 2.2 GHz CPU (2 flops per clock)
  - 4.4 Gflops peak, 620 Mflop sustained (14% of peak)
- Peak [M,G,T,P] flops is not a good indicator of performance for 'real' codes
  - Was ever so for vector v scalar
  - Also true for scalar CPUs
- P690+ achieves this as a 32-way SMP
  - More general purpose architecture (not all codes are MPI parallel)
  - Permits the use of OpenMP
  - For IFS this results in better scalability for large number of nodes

RAPS8 IFS 4D-Var T_L511L60 using 40 CPUs



RAPS8 IFS 4D-Var T_L511L60 using 52 CPUs



IFS T_L511L60 4D-Var

- T_L511L60 4D-Var is ECMWF's operational resolution
  - Problem is too big for our small Opteron cluster
  - Should run in < 1 hour
  - Validation of RAPS8 benchmark on non-IBM system
- Efficient parallel filesystem is essential for 4D-Var
  - P690+ GPFS / Federation ~ 1300 MB/s (2176 CPUs)
  - Opteron cluster Lustre / IP over Gigabit Ethernet ~ 100 MB/s (64 CPUs)
  - Performance of both filesystems are scalable
- IFS 4D-Var and Model show similar relative performance against IBM P690+ (IBM 8% faster)

Compilers for AMD Opteron

- Experience with Portland pgf90 v5.1 disappointing
  - 6 bug reports compiling RAPS7_fc IFS
  - No response for Portland Group
- pgf90 v5.2
  - Full F95 support
- Absoft Fortran95 v9.0
  - Reliable compiler, but slowest generated code
- PathScale pathf90 v1.2 (used for IFS runs)
  - Both reliable and good performance
  - Only one routine resulted in a runtime problem at high optimization (-O2)
  - Support for reading IBM unformatted (big endian) files

Other S/W

- System administration - Clusterworx (Linux Networx)
- Batch subsystem - Sun Grid Engine (being integrated)
- System monitoring - GANGLIA
- Debuggers
  - Now: print,*
  - Future: Totalview / DDT
- Parallel file system - Lustre 1.2.3 (CFS)
  - 6 nodes dedicated for Lustre (1 MDS, 5 OSS)
  - Raw write perf to single RAID5 device (FAStT600) ~ 35 MB/sec
  - With Lustre ~15 MB/sec
  - Today: GigaBit Ethernet 120 MB/sec, Lustre aggregate write ~ 90 MB/sec
  - Future: Lustre over Infiniband

Conclusions

- Performance of AMD Opteron / Infiniband is very competitive
  - Infiniband bandwidth limited by PCI-X today
  - Binding tasks to CPUs on Opteron (future)
- Support model when S/W comes from several sources?
- Software infrastructure sufficient to deploy cluster as a general purpose server
- Scalability to large o(1000) CPU cluster
  - Larger switches needed
  - Reliability
- Need experience first with use as a GP server

RAPS8 IFS 4D-Var T_L511L60 on IBM P690+