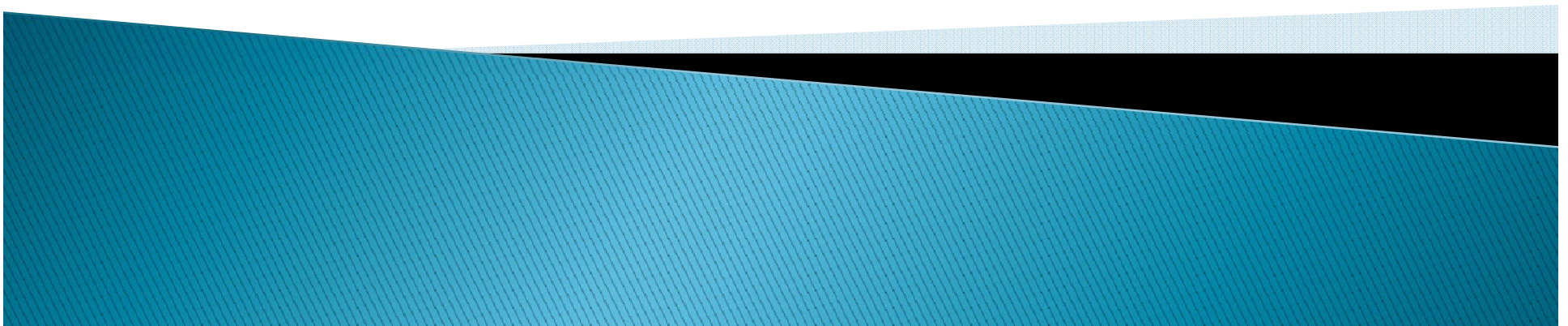
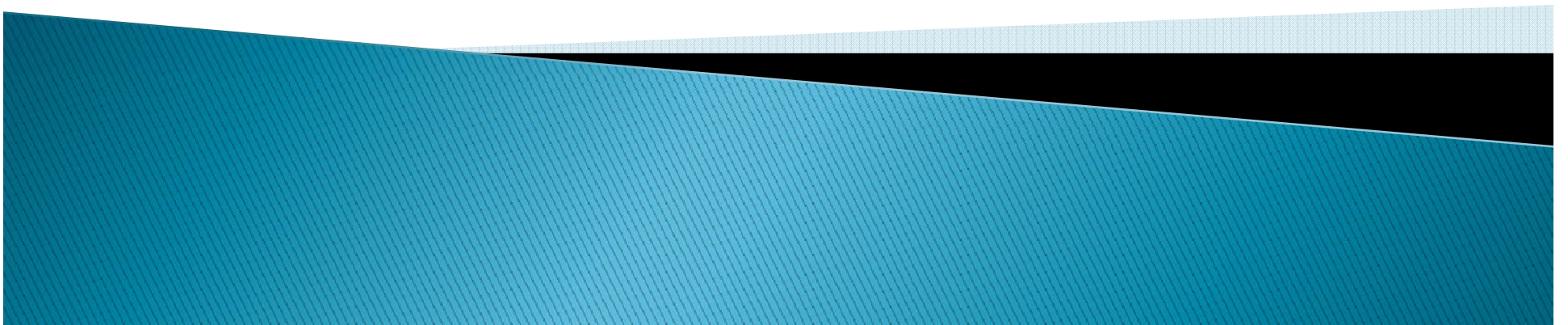


SHPCP InfiniBand Benchmark Results

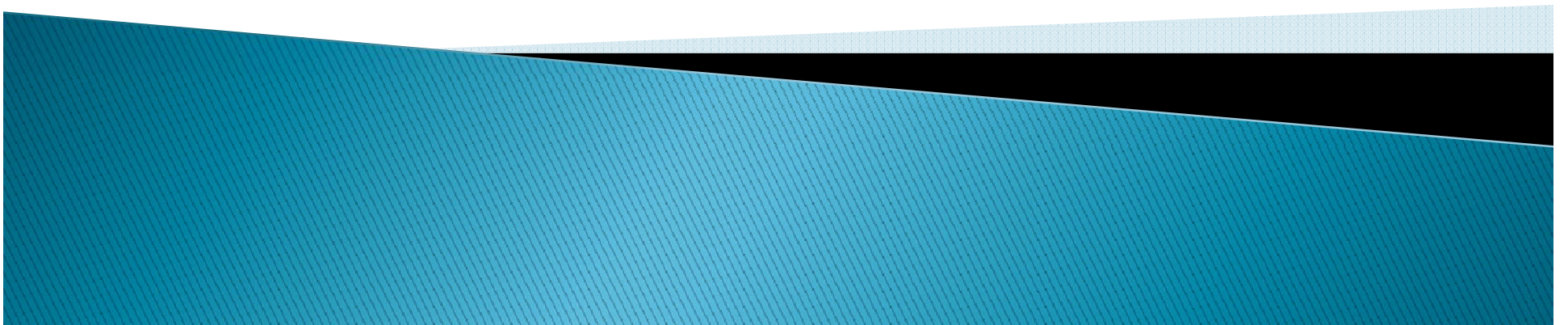
November 3, 2010 Technical Meeting



**Thank You:
QLogic Corporation**



**Thank You:
Max Dechantsreiter**



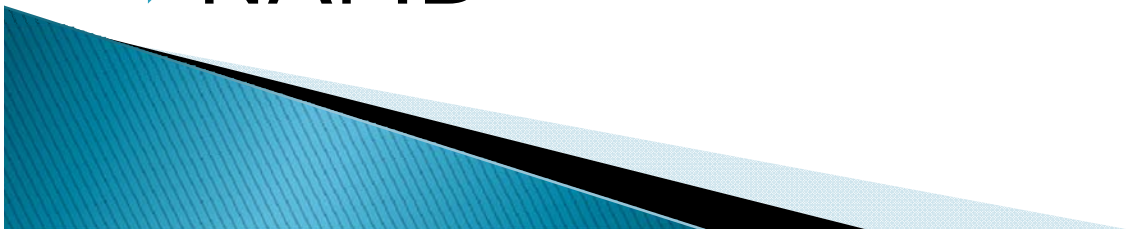
Applications Benchmarked

- ▶ IWAVE

- ▶ WRF

- ▶ IQCS

- ▶ NAMMD



Applications Benchmarked

▶ IWAVE

▶ CWP/SU: Seismic Un*x Release 42:
▶ <http://www.cwp.mines.edu/cwpcodes/>
▶ (built with GNU)

▶ IWAVE:
▶ <http://www.trip.caam.rice.edu/software/iwave/doc/html/index.html>
▶ (Thanks to William W. Symes, Rice University)

▶ WRF

▶ NetCDF 4.1.1:
▶ <http://www.unidata.ucar.edu/software/netcdf/>

▶ WRF Version 3.2:
▶ <http://www.mmm.ucar.edu/wrf/users/wrfv3.2/updates-3.2.html>
▶ 12km CONUS [Continental United States] benchmark:
▶ http://www.mmm.ucar.edu/WG2bench/conus12km_data_v3-2/
▶ (Thanks to John Michalakes, NREL)

▶ Note: the "standard" WRF benchmark:
▶ <http://www.mmm.ucar.edu/WG2bench/>
▶ works only with WRFV3.1.1 and earlier;
▶ WRFV3.2 runs slower than WRFV3.1.1, so results cannot be compared

▶ IQCS

DEISA benchmarks:

▶ <http://www.deisa.eu/science/benchmarking>
▶ g (Distributed European Infrastructure for Supercomputing Applications)

▶ IQCS (Improving Quantum Computer Simulations)
▶ "[...] serves primarily as a benchmark for memory bandwidth and internode communication."

▶ <http://www.deisa.eu/science/benchmarking/codes/iqcs>

▶ NAMD

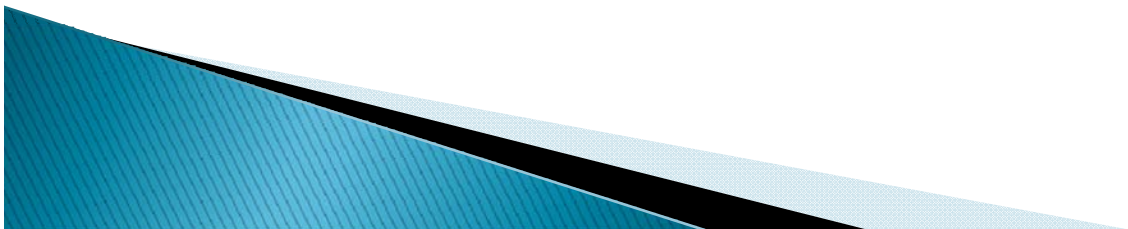
▶ NAMD 2.7
▶ <http://www.ks.uiuc.edu/Research/namd/>

▶ ApoA1 benchmark: 92,224 atoms, 12A cutoff + PME every 4 steps, periodic

▶ <http://www.ks.uiuc.edu/Research/namd/performance.html>

Environment

- ▶ EVERYTHING aside from SU was built with:
 - Intel 11.1.059 compiler
 - Intel MPI 4.0.0.025
- ▶ EVERYTHING was "out of the box"
 - NO OPTIMIZATIONS
 - Some decompositions selected in order to minimize load imbalance
- ▶ QLogic InfiniBand QDR
 - 1:1 subscription (oversubscription to follow after meeting)
 - No 10GE testing (data to follow after meeting)
- ▶ IBM iDpx 360M3 (Westmere 2.93 GHz)



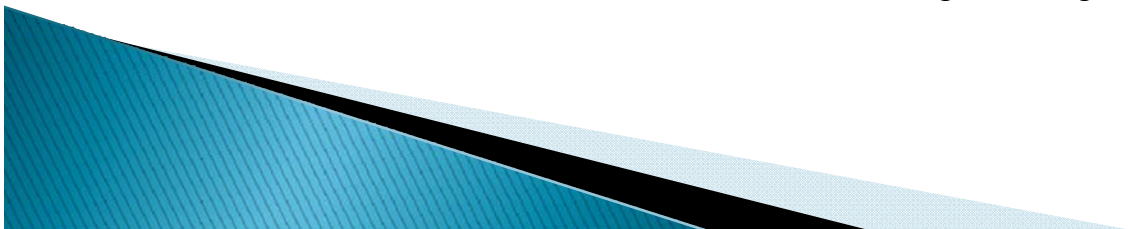
InfiniBand Set Up

- ▶ No Oversubscription (1:1)
- ▶ I_MPI_DEVICE=shm:X
 - X = tcp
 - X = dapl
 - X = tmi
- ▶ QDR
- ▶ Native (true) TCP/IP

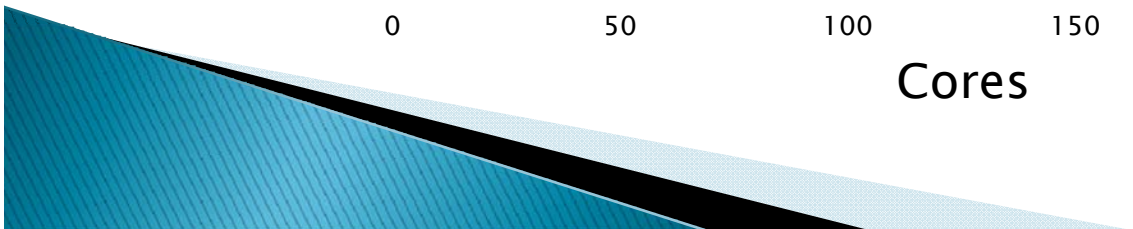
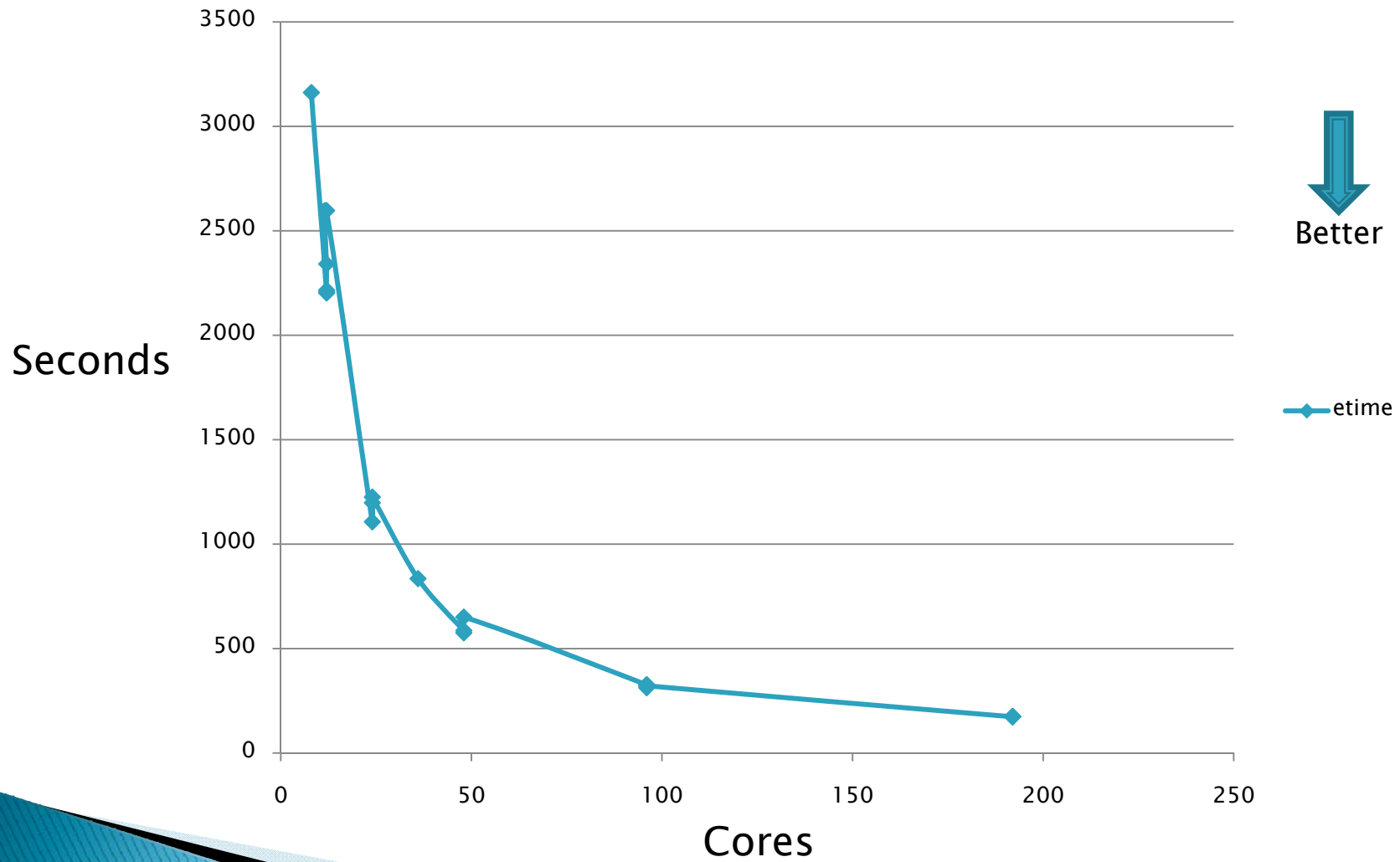


Heading Definitions

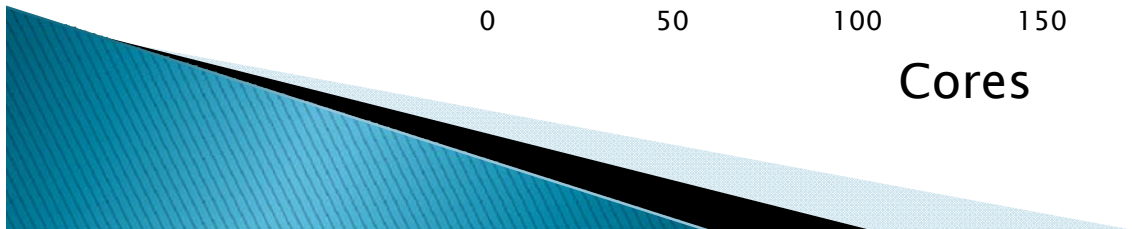
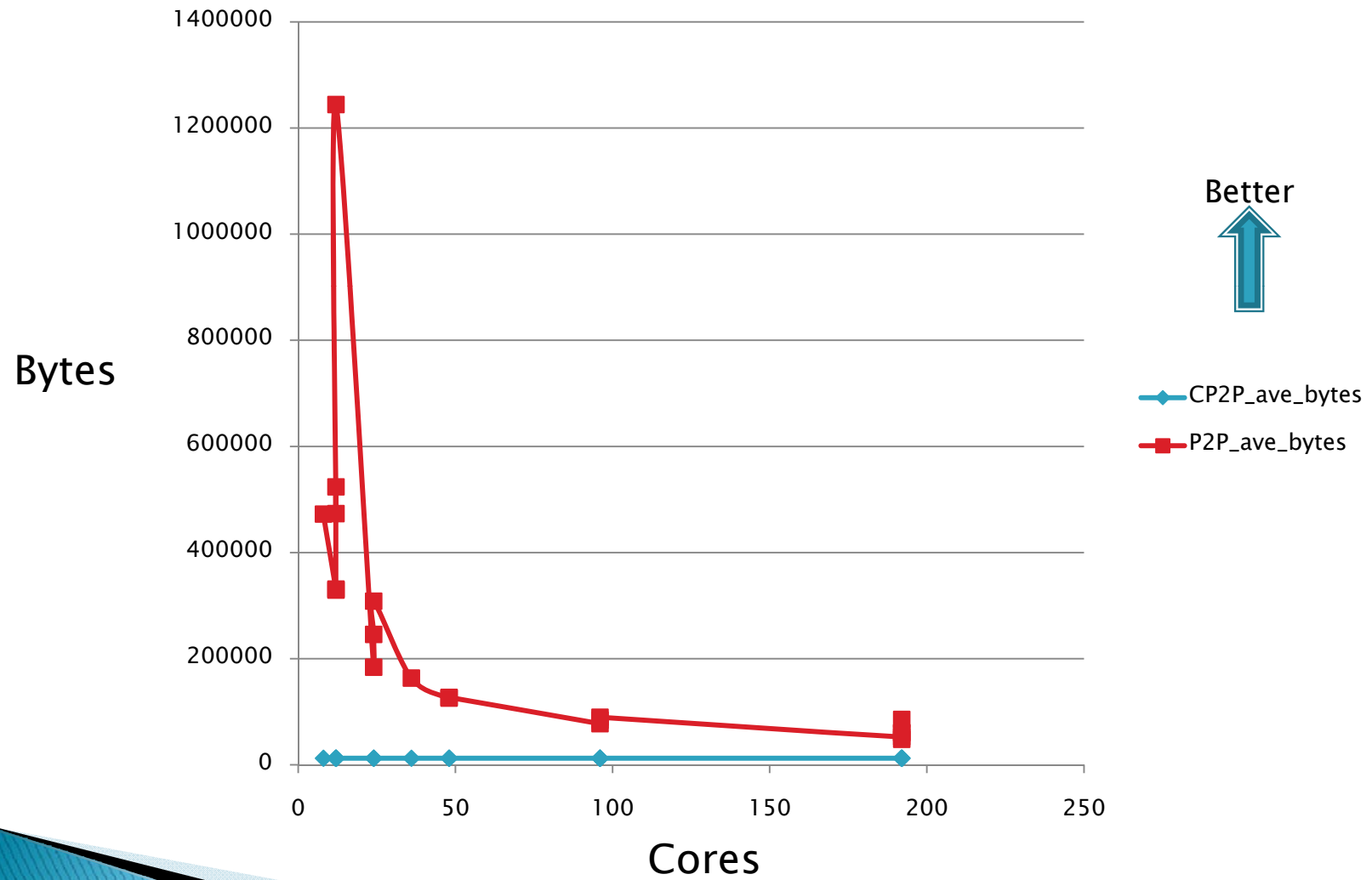
cores	number of CPUs or cores per chip
decomp	decomposition strategy or topology
tpn	tasks per node (synonymous with ppn = processes per node)
NBITS	number of qubits or states
sim_speed	Simulation speed is the model time step, 72 seconds, divided by average time per time step. Gigaflops per second is simulation speed times 0.418 for this case based on a measured operation count of 30.1 billion floating point operations per second (or simply the operation count divided by the average time per time step).
GB_total	aggregate transfer rate in GB/s
GB_intra	intranode transfer rate in GB/s
GB_inter	internode transfer rate in GB/s
GB_proc_max	max. average transfer rate (GB/s) by any one process
CP2P_ave_bytes	average point-to-point message length (in bytes) within collective operations (Bcast, Allreduce, etc.)
P2P_ave_bytes	average point-to-point message length (in bytes) outside of collective operations (Send, Bsend, etc.)
etime	elapsed time in seconds
intranode transfer	Sum of all the amounts sent between tasks running on the same nodes to get the upper limit on intranode transfers: this is what the amount of data transferred via shm would be, IF all the messages were shorter than the threshold for shm
internode transfer	Sum of volume of data transferred internode could be larger, if there were intranode messages too long to use the shm buffers: these messages would go out on the network - although perhaps by loopback.



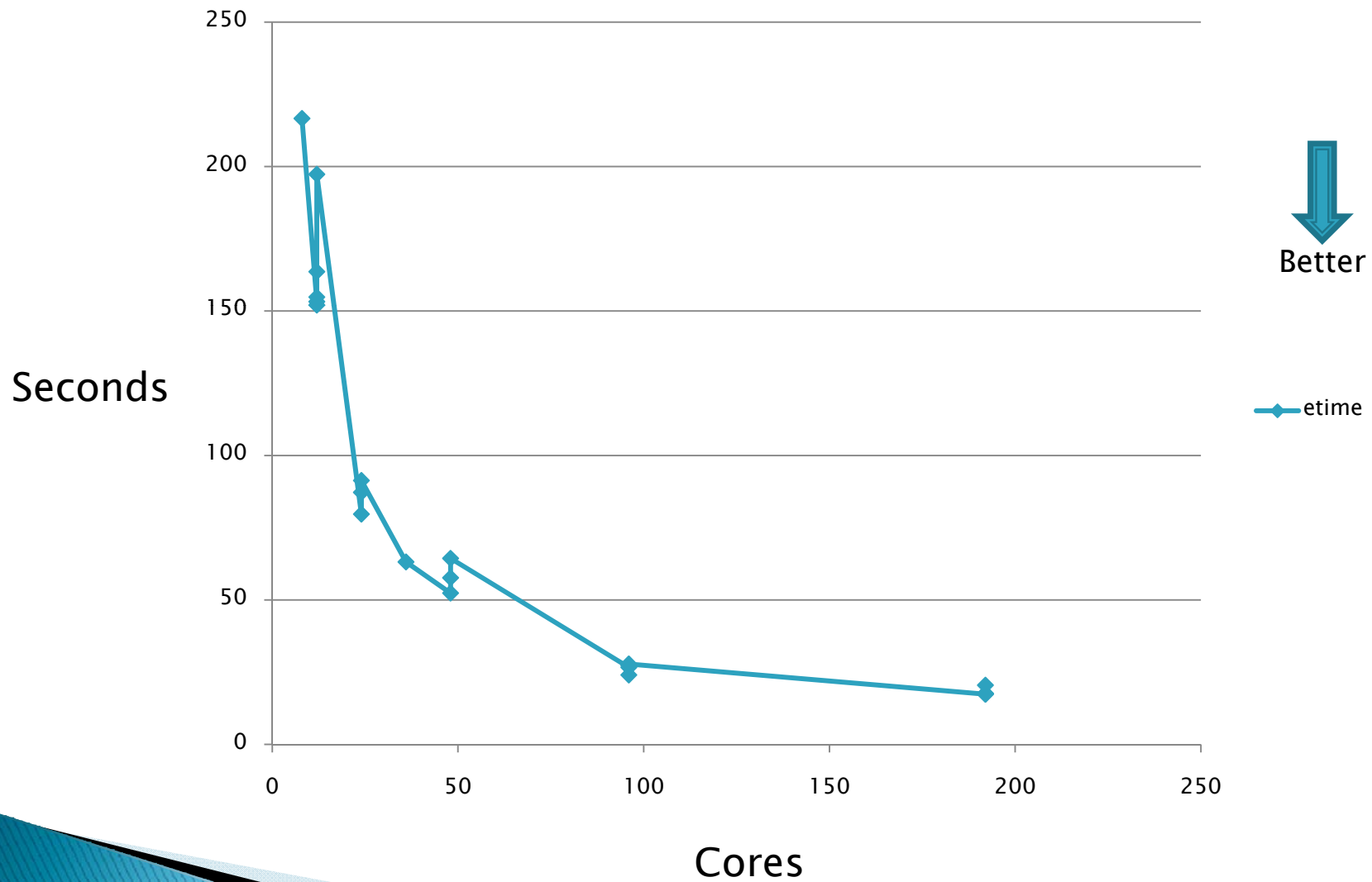
IWAVE 10m



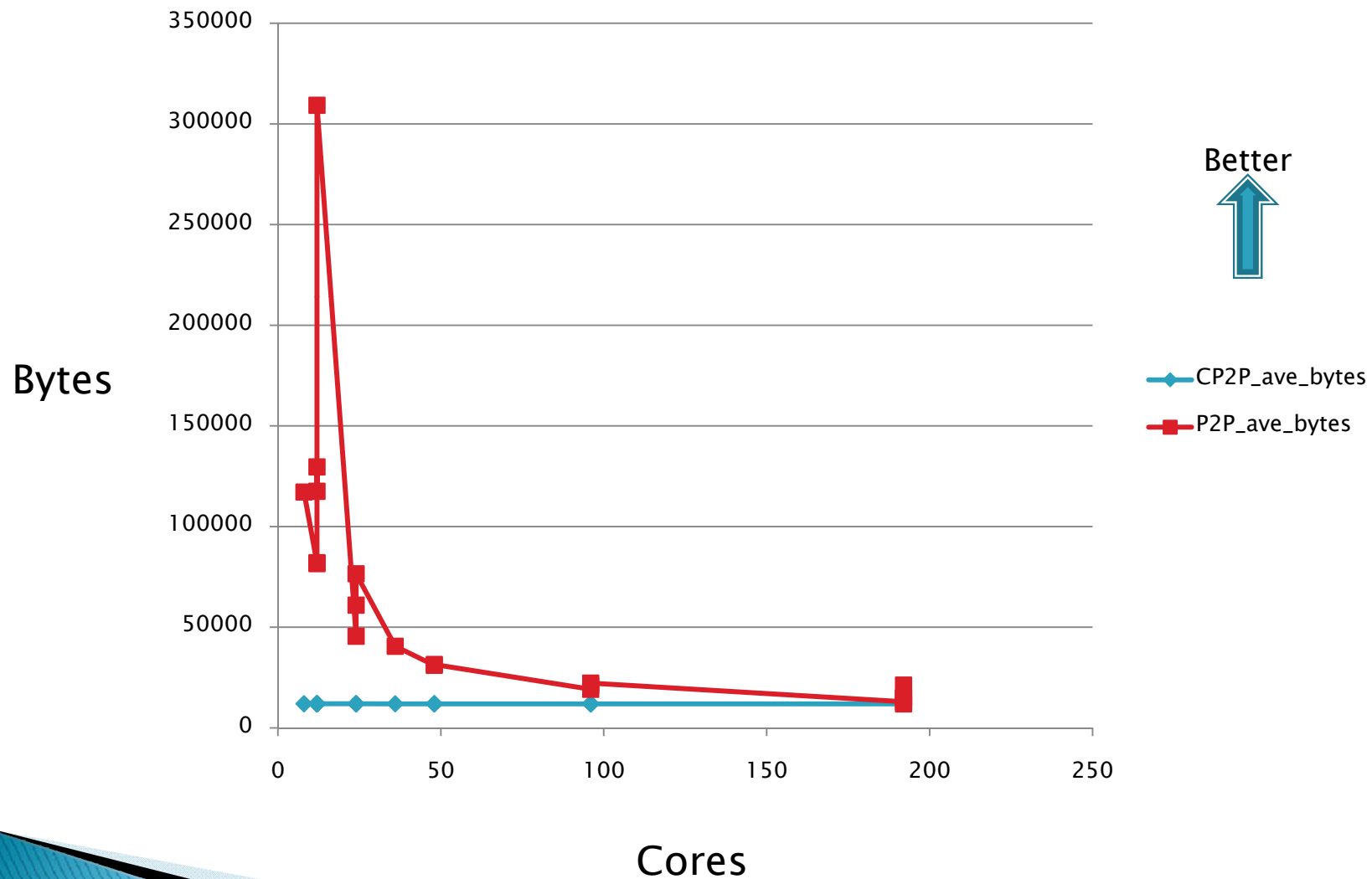
IWAVE 10m



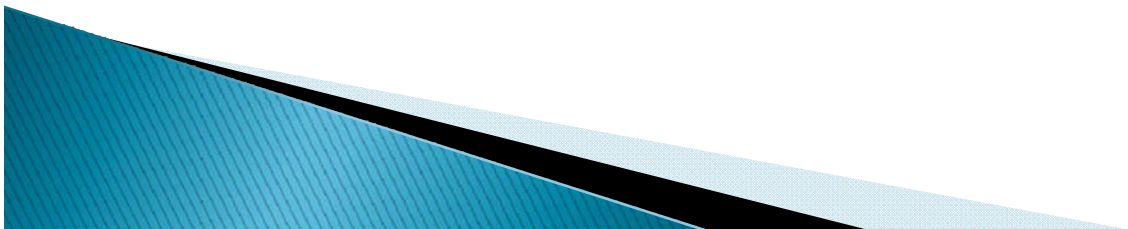
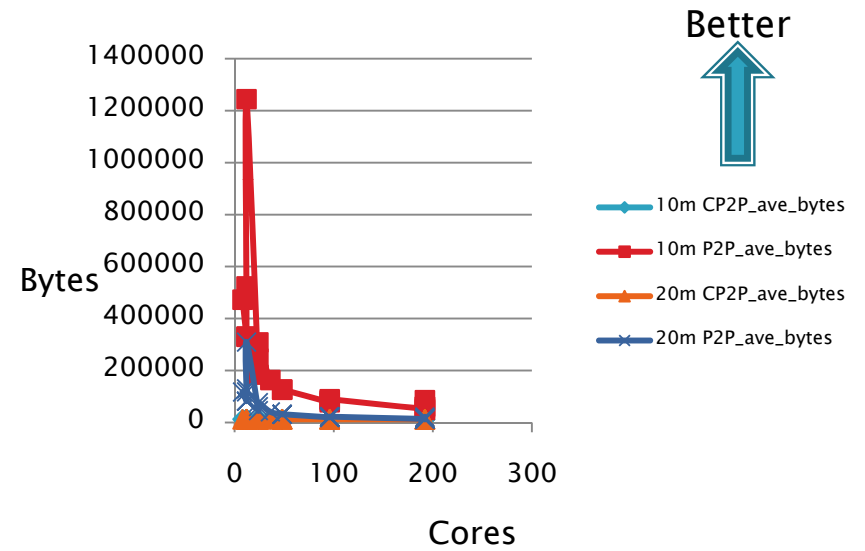
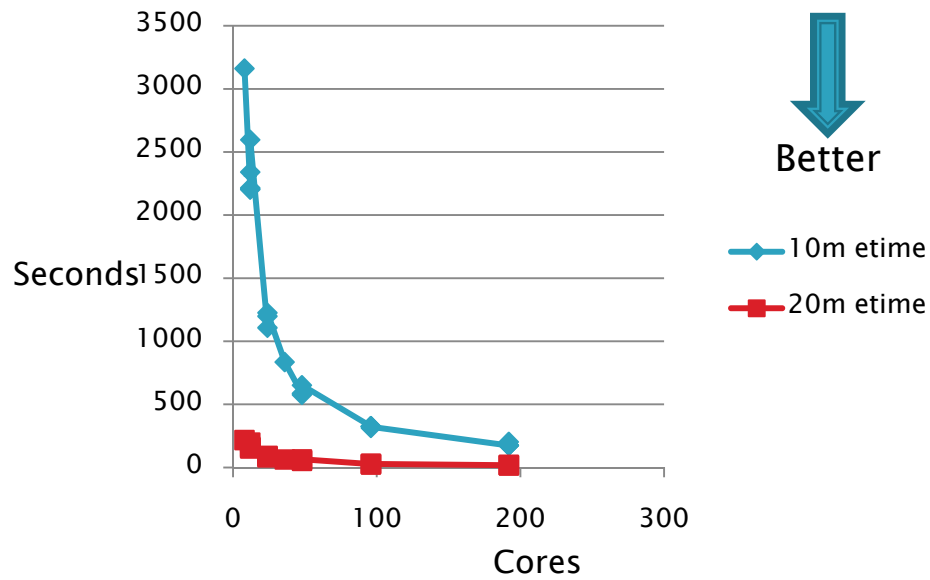
IWAVE 20m



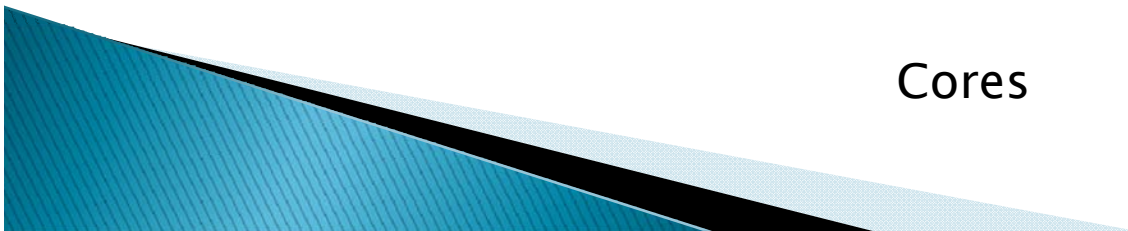
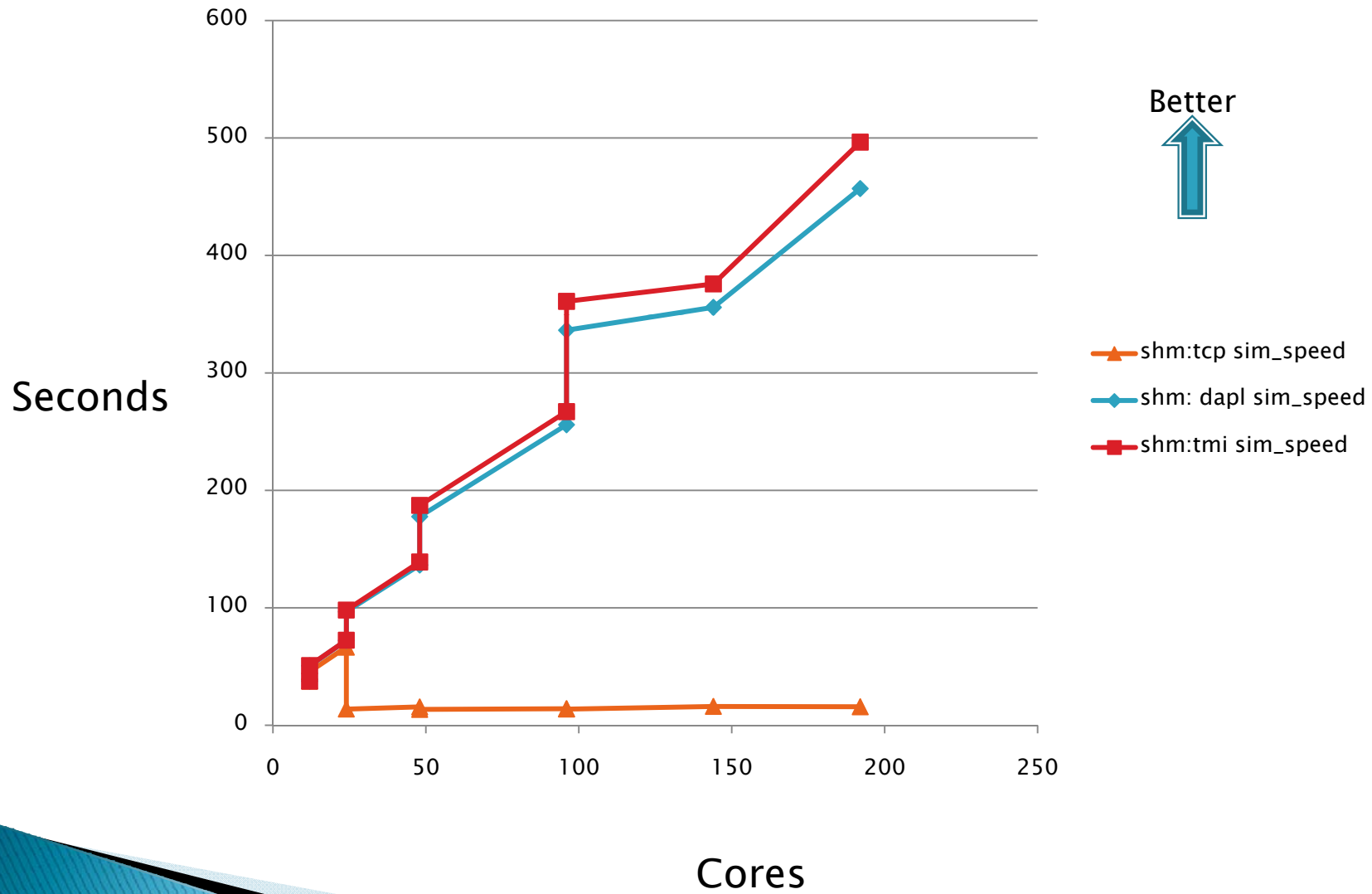
IWAVE 20m



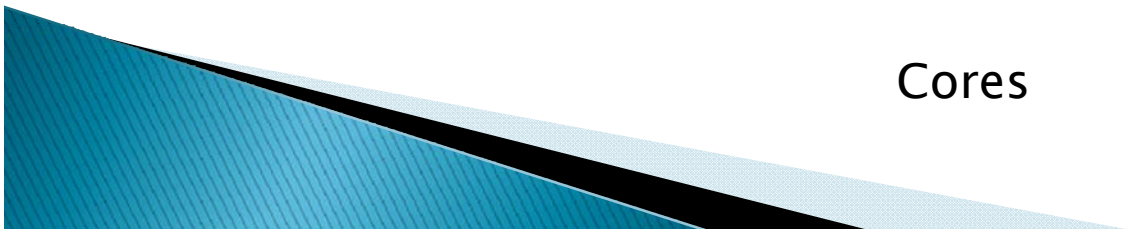
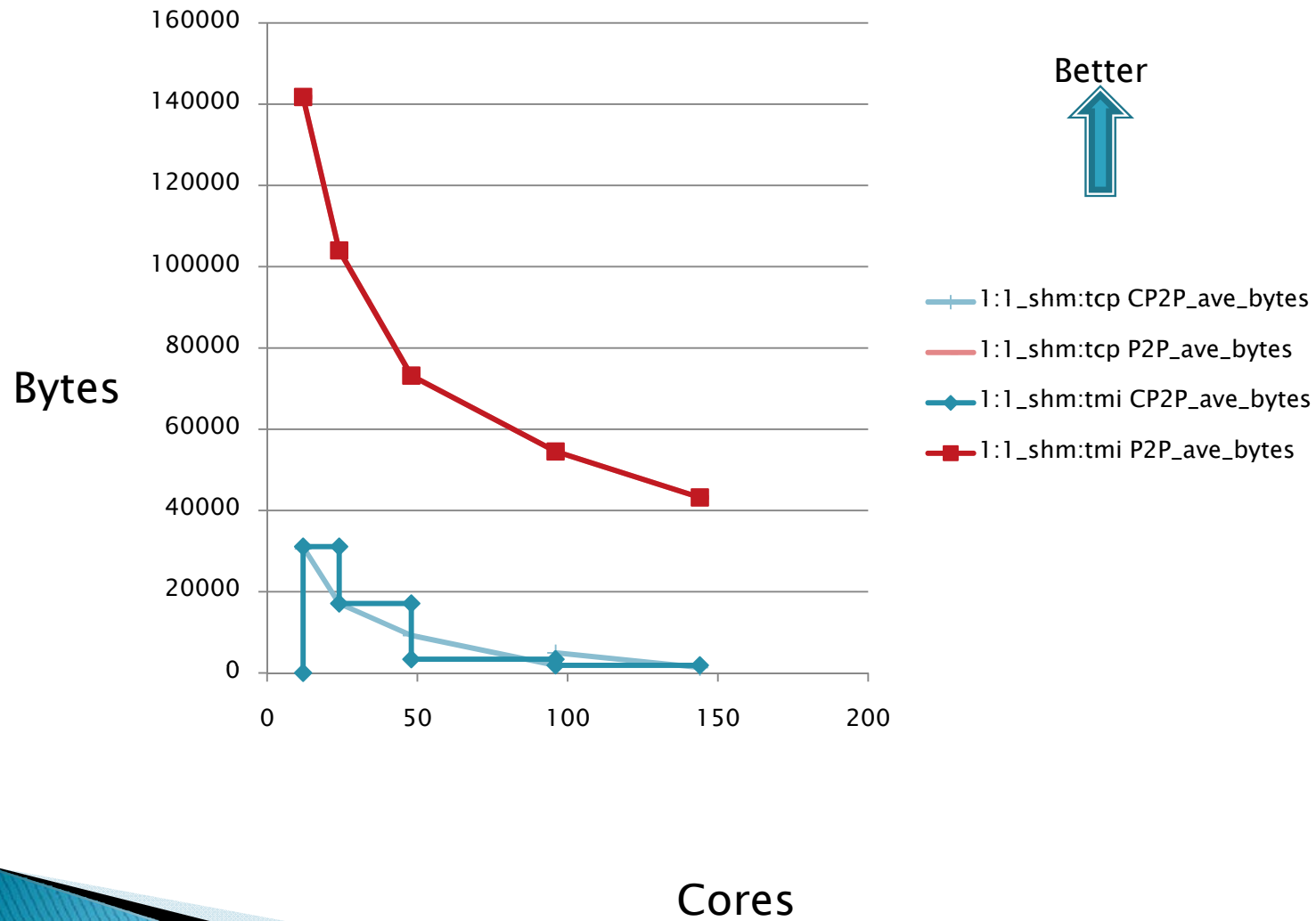
IWAVE 10m vs.20m



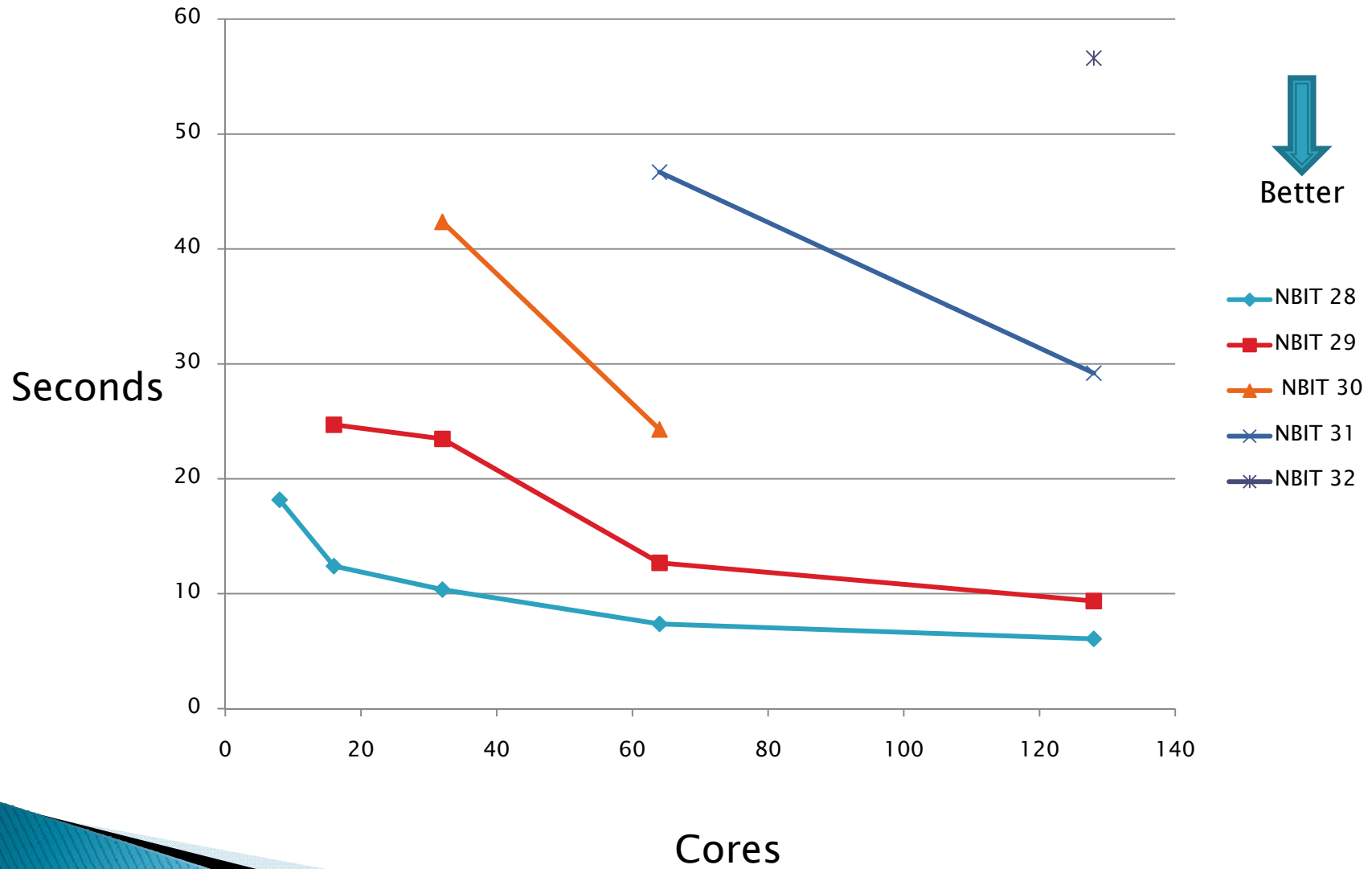
WRF



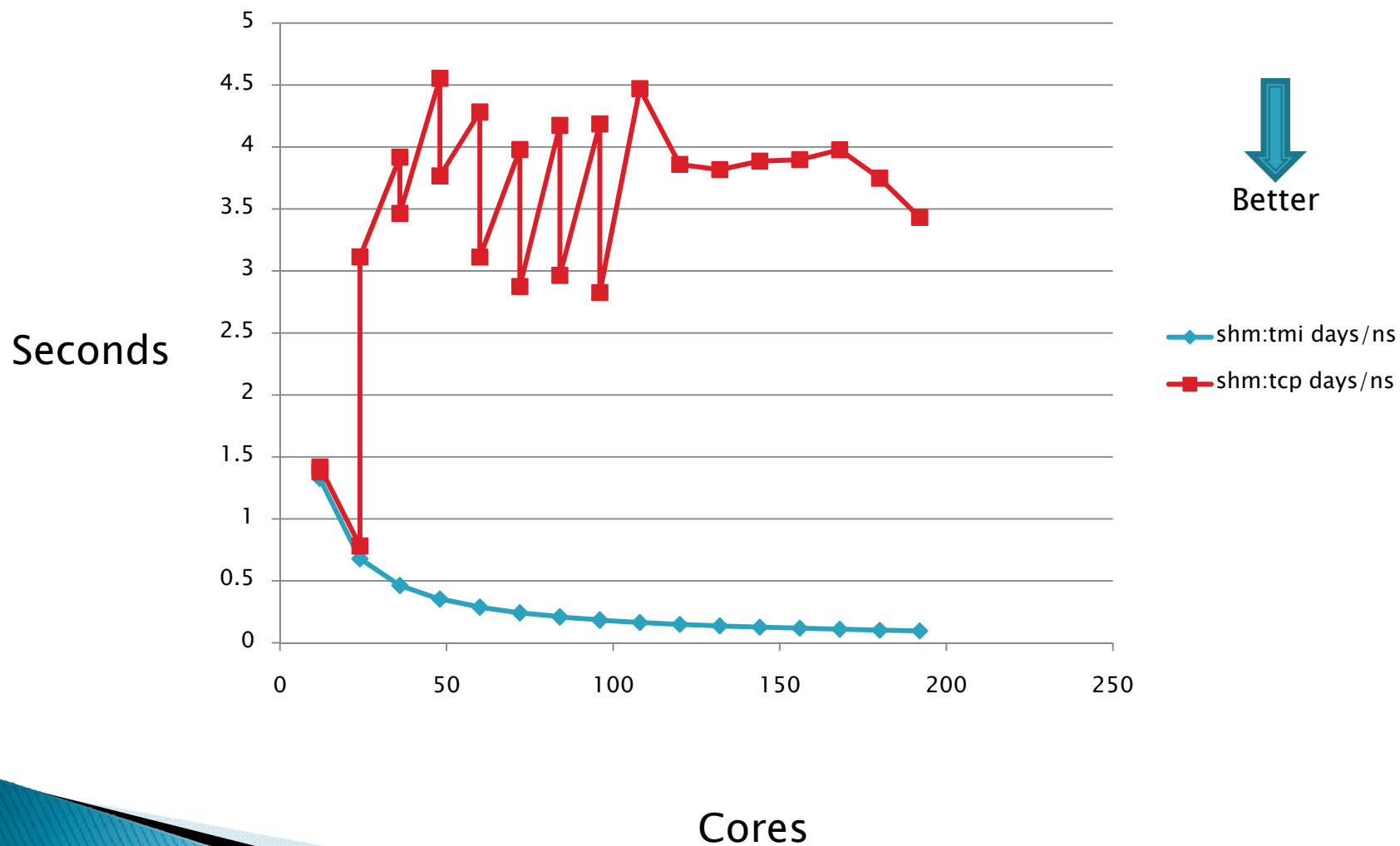
WRF



IQCS



NAMD



NAMD

