Whole System Performance Instrumentation

(would be very useful in differentiating what are application-specific problems and what are not, and to identify system performance problems directly, and immediately, rather than implicitly, postmortem)

Patrick H. Worley
Oak Ridge National Laboratory

The Salishan Conference on High-Speed Computing
April 23, 2008
Salishan Lodge
Gleneden Beach, Oregon

Oak Ridge National Laboratory

Office of Science
U.S. DEPARTMENT OF ENERGY

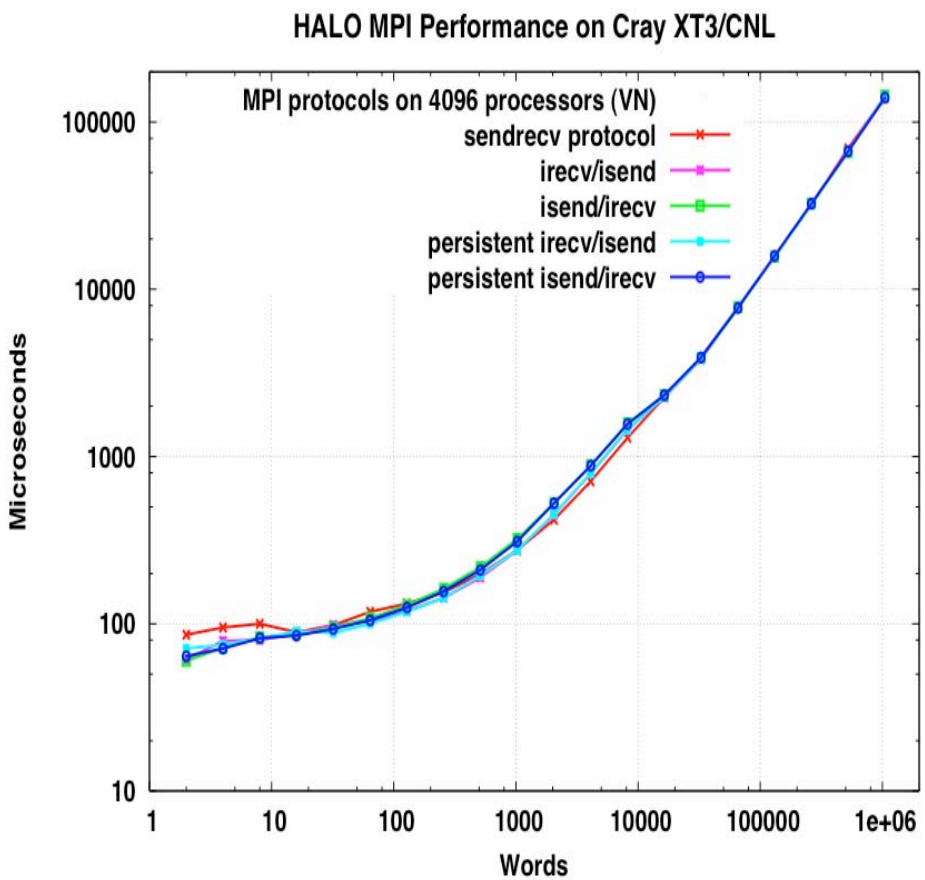
Acknowledgements

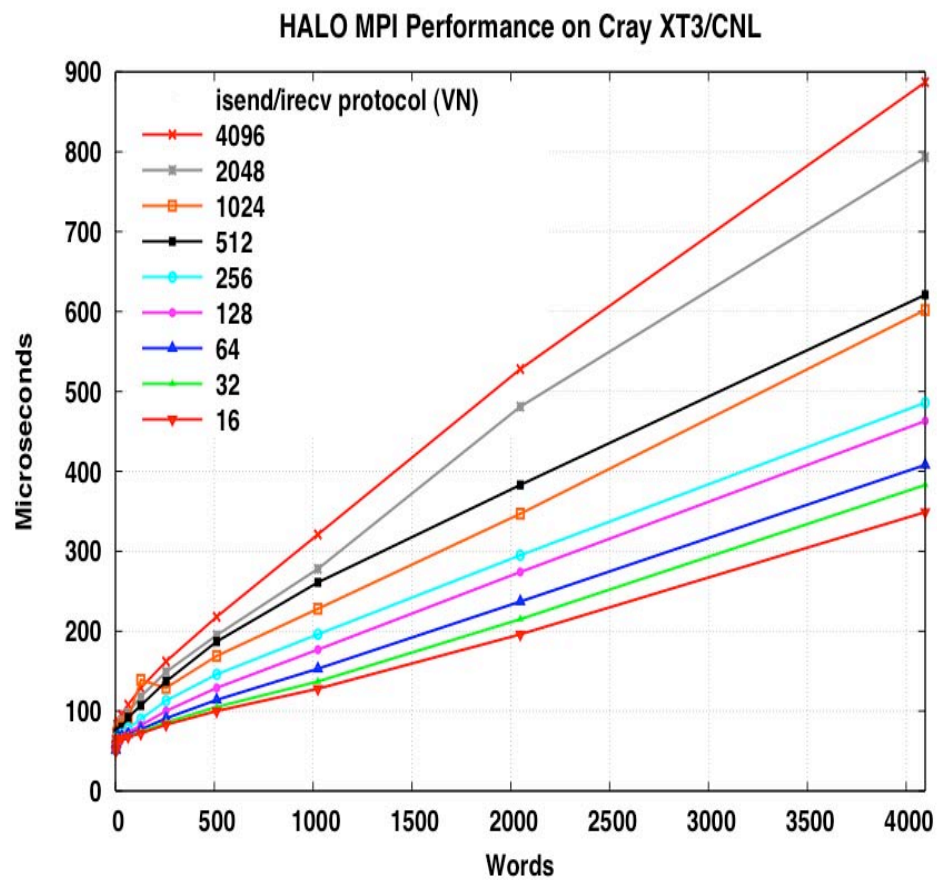
- The work described in this presentation was sponsored by the Office of Mathematical, Information, and Computational Sciences, Office of Science, U.S. Department of Energy, under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.
- These slides have been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725 with UT-Battelle, LLC. 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.
- This work used resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the Department of Energy under Contract DE-AC05-00OR22725, and of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Overview

- First half of following slides were presented at Cray Technical Workshop on February 27, 2008, describing results collected the previous weekend. The issue is significantly different performance on two very similar systems, the Cray XT3 at ORNL (**jaguarcnl**) and the Cray XT4 at NERSC (**Franklin**). My conjecture was that the other users were impacting performance of these microkernels on Franklin, but not on jaguarcnl, but how can I tell?
- My current conjecture is that performance anomalies of this sort will become even more prevalent at scale, and engineering-in (improved) support for whole system performance "health" monitoring would be a service to both the users and the system administrators.
- In the meantime, I have begun adding support for (low overhead, basic) application profiling that is dumped periodically to be monitored via workflow systems such as the Dashboard being developed by S. Klasky et al as part of CPES.

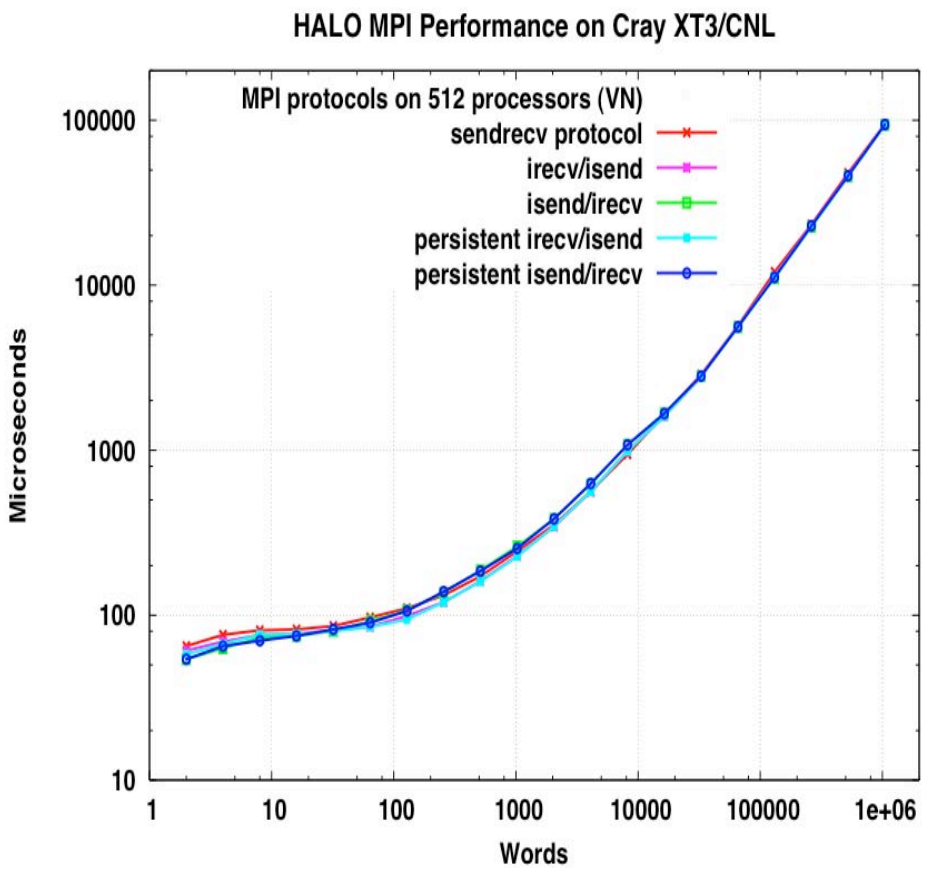
HALO on Jaguarcnl

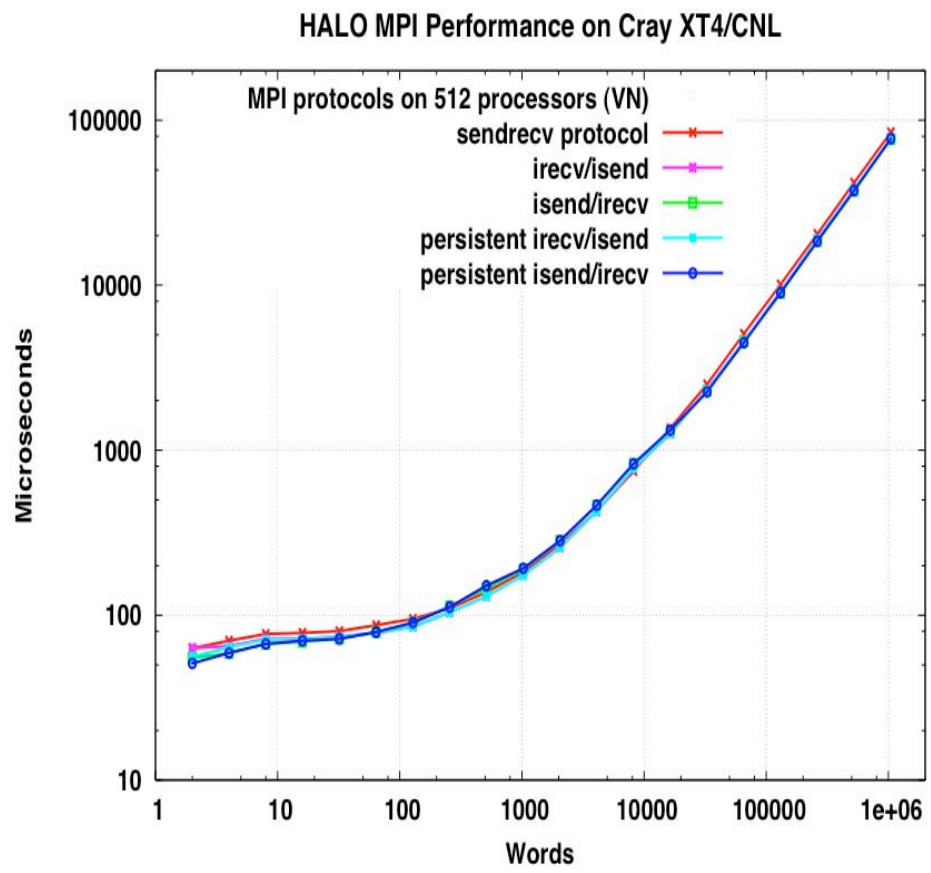




Alan Wallcraft's HALO benchmark: Measures max time (over all processes) for a small number of repetitions, normalized by number of repetitions. Little protocol sensitivity is evident. Cost does grow with process count (increasing by a factor of 3 between 16 and 4096 processes). This does not explain source of increase in cost.

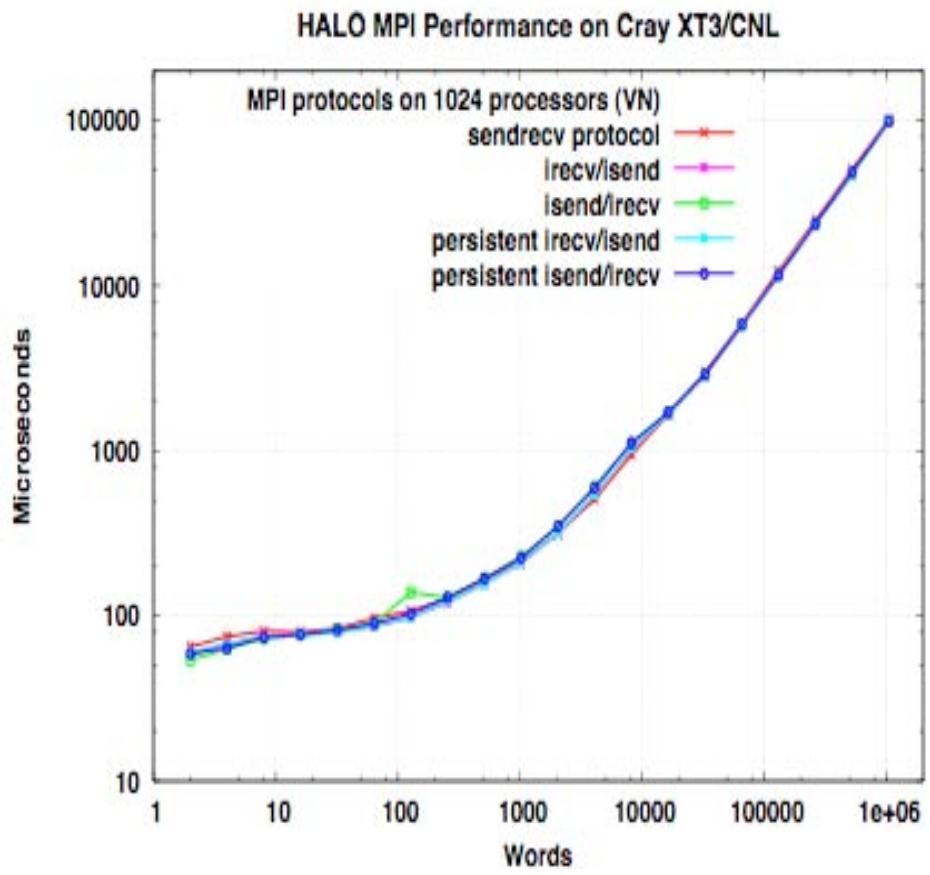
HALO on Jaguarcnl vs. Franklin (512)

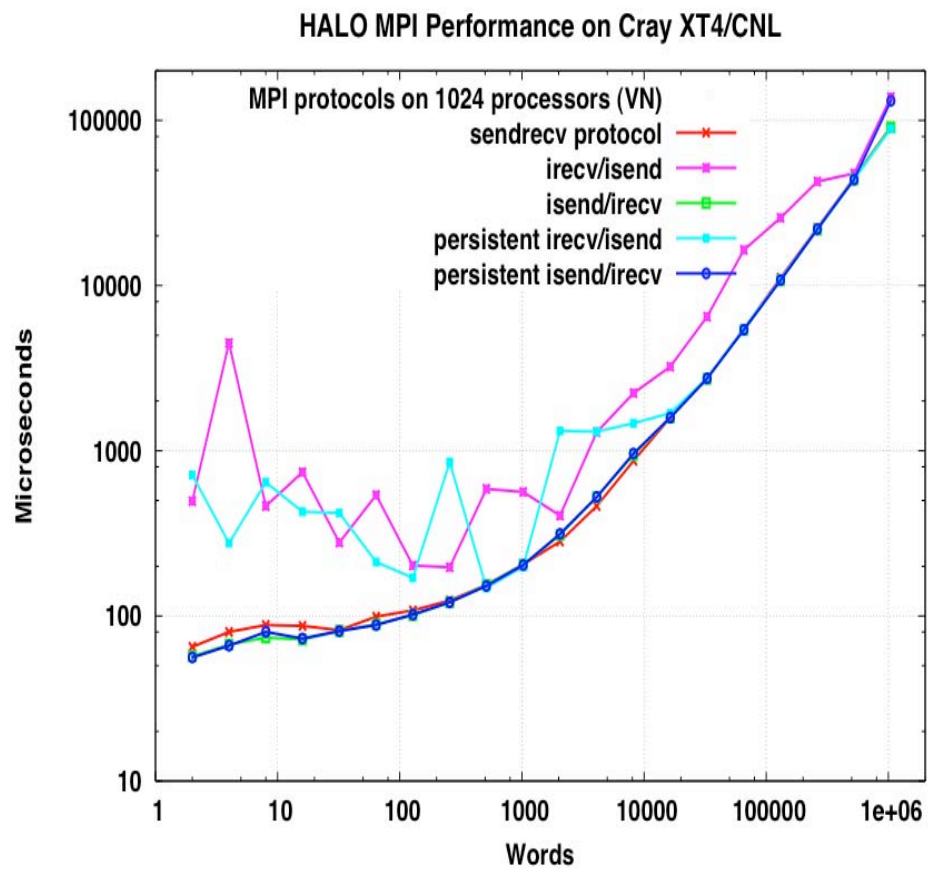




HALO benchmark shows similar behavior on both Franklin and Jaguarcnl for this experiment.

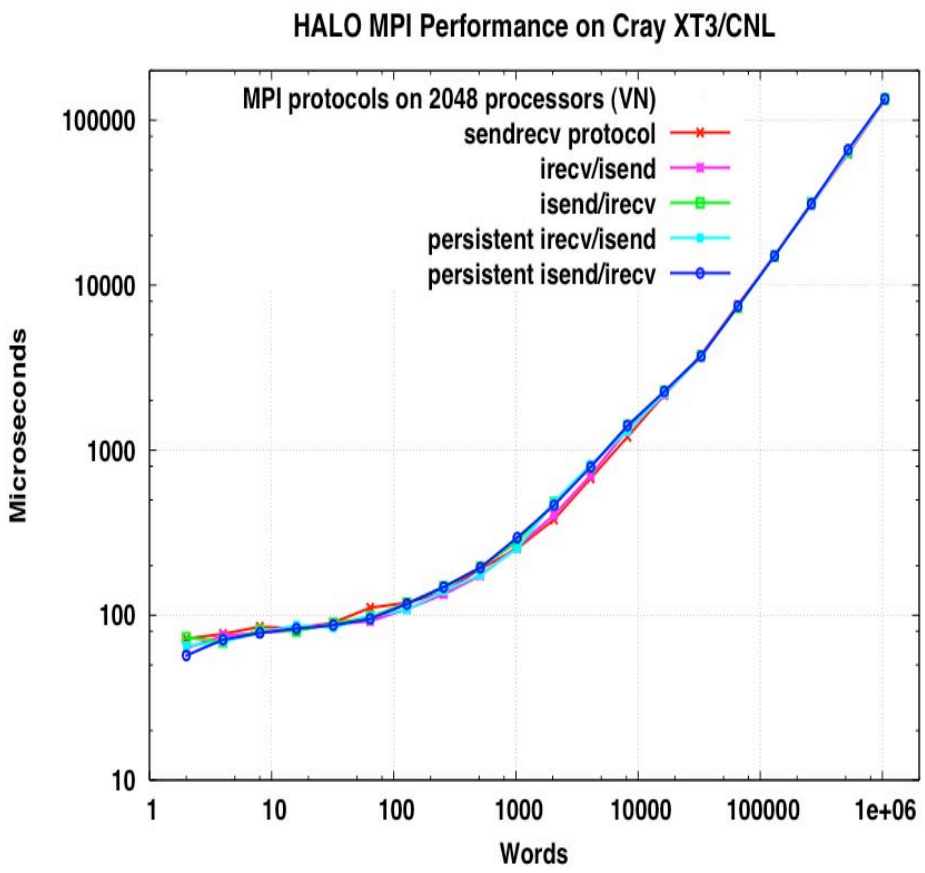
HALO on Jaguarcnl vs. Franklin (1024)

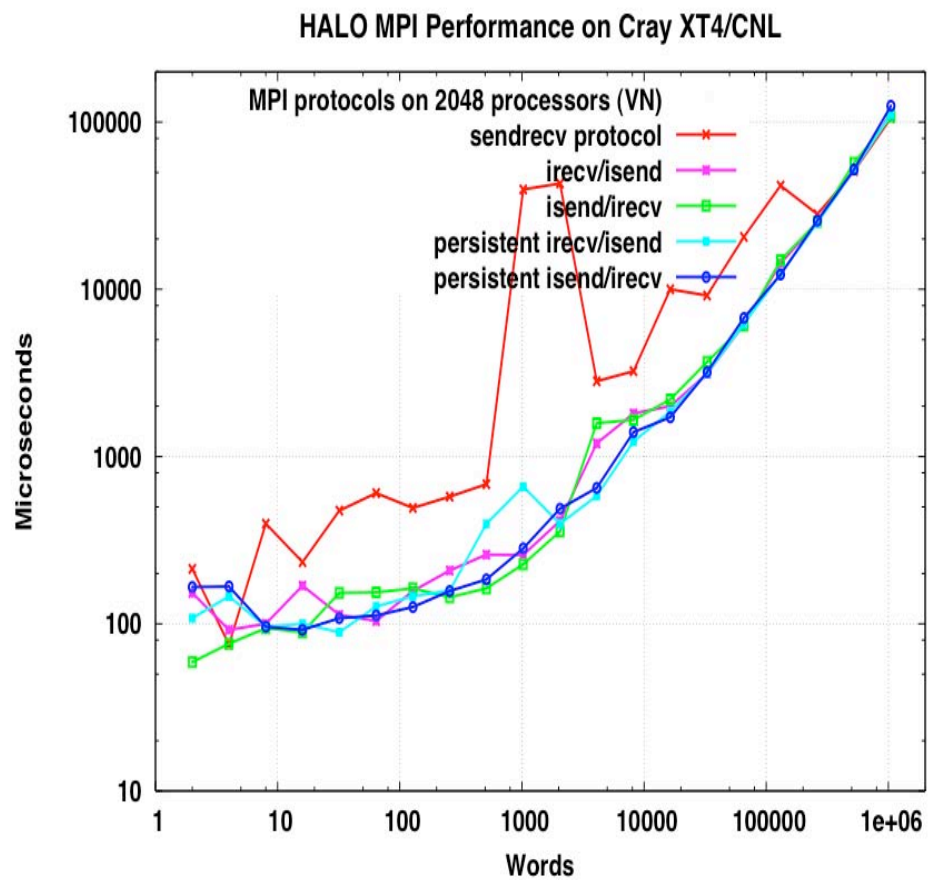




Some experiments on Franklin show performance variability that did not see on Jaguarcnl over the same period of time. Difference in workload?

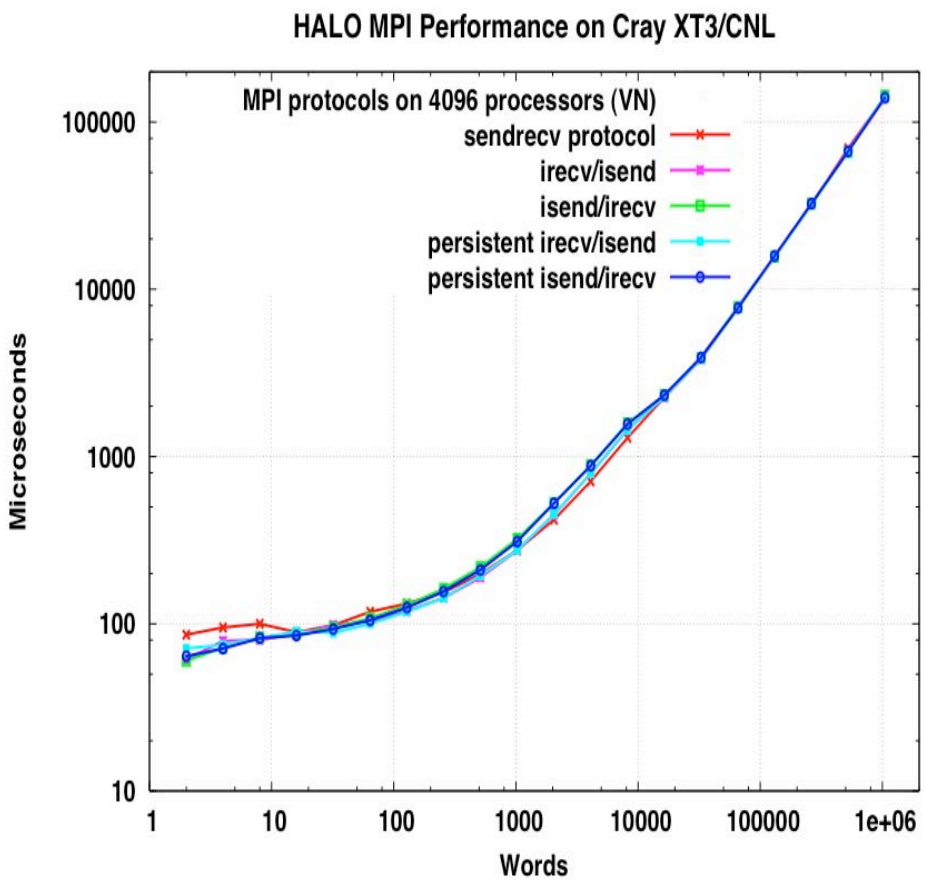
HALO on Jaguarcnl vs. Franklin (2048)

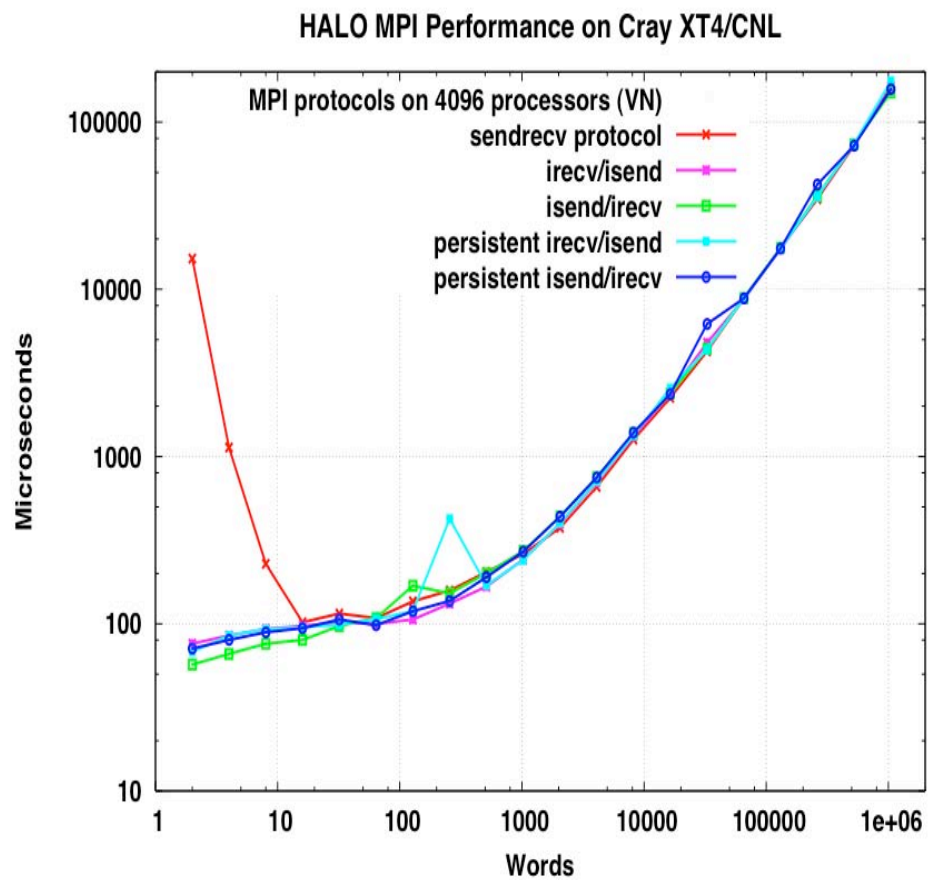




Variability on Franklin does not appear to be a function of MPI protocol.

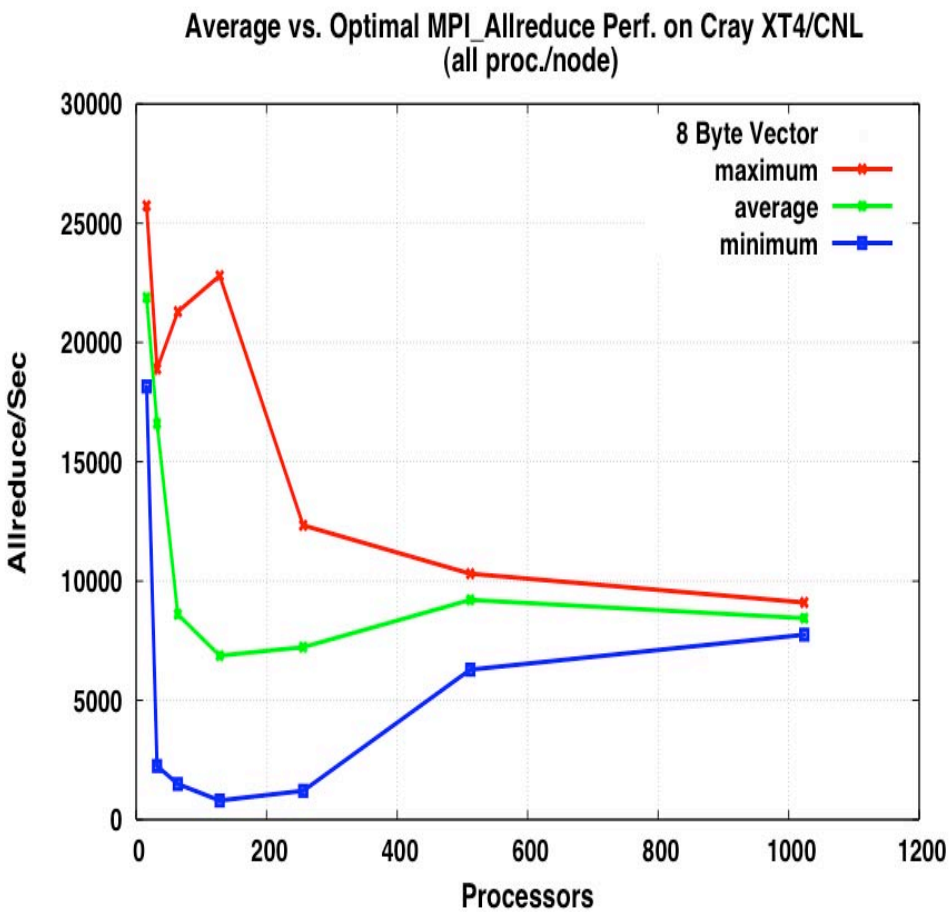
HALO on Jaguarcnl vs. Franklin (4096)

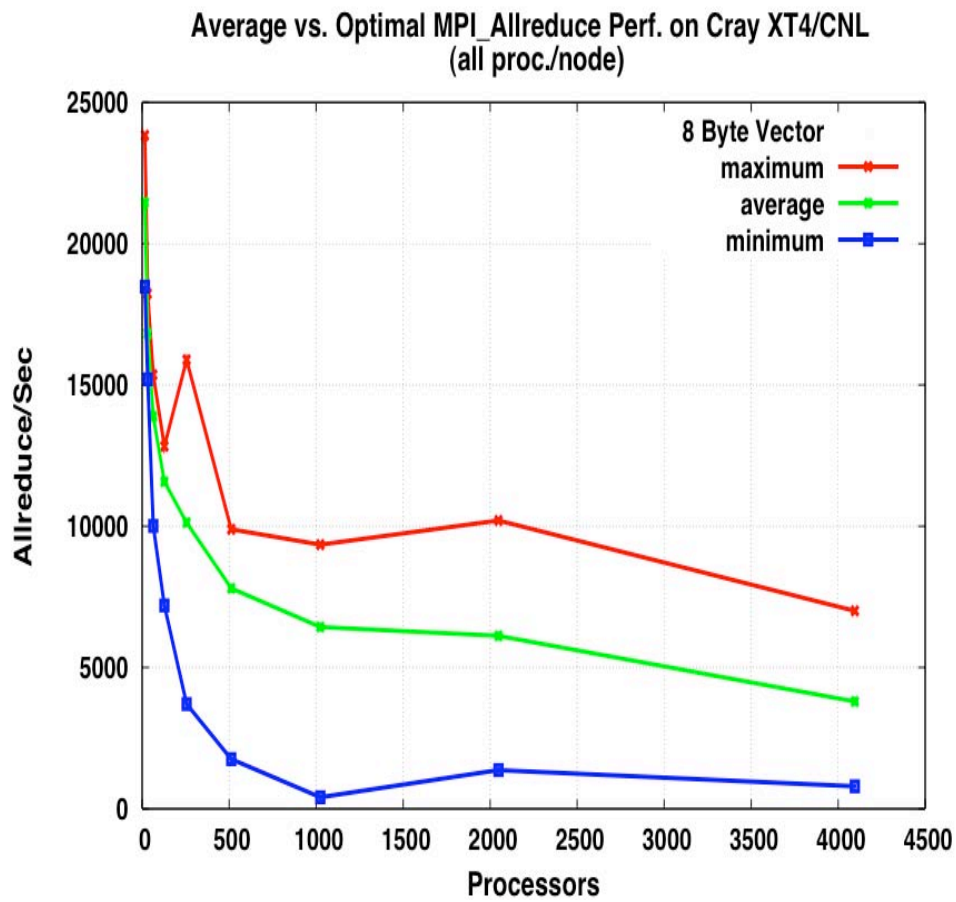




Variability diminishing as go to higher processor counts (essentially gone for 8192 processes) - coincidental? Note that performance variability on Franklin similar for SN mode experiments.

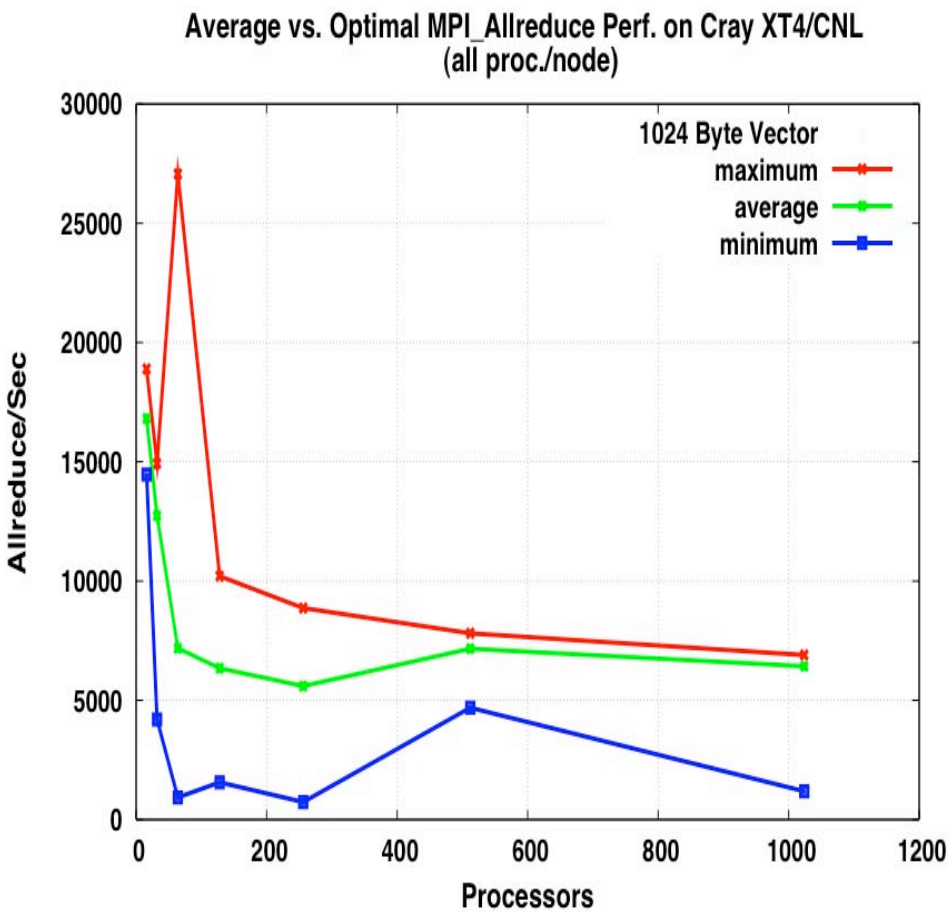
Allreduce "jitter" test on Franklin (8B)

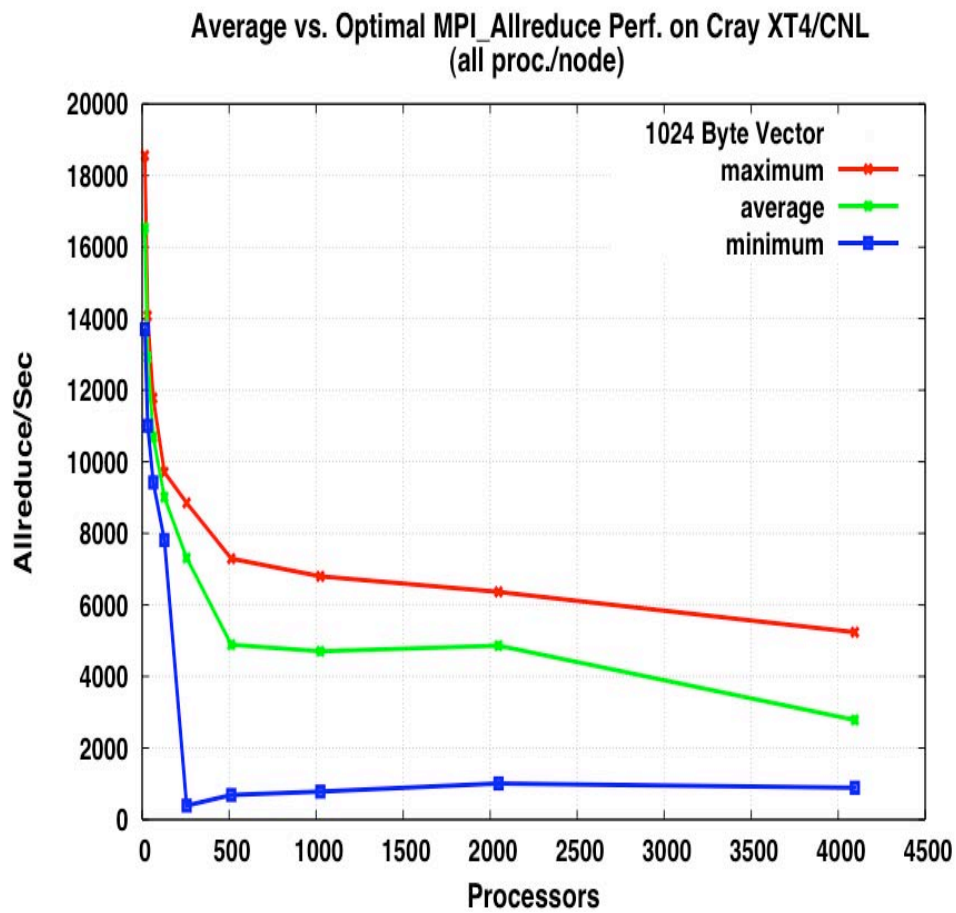




Experiments on two different days (separate runs for each processor counts). Perturbations not the same each day, but worst and average cases (80 samples) much worse than best observed performance.

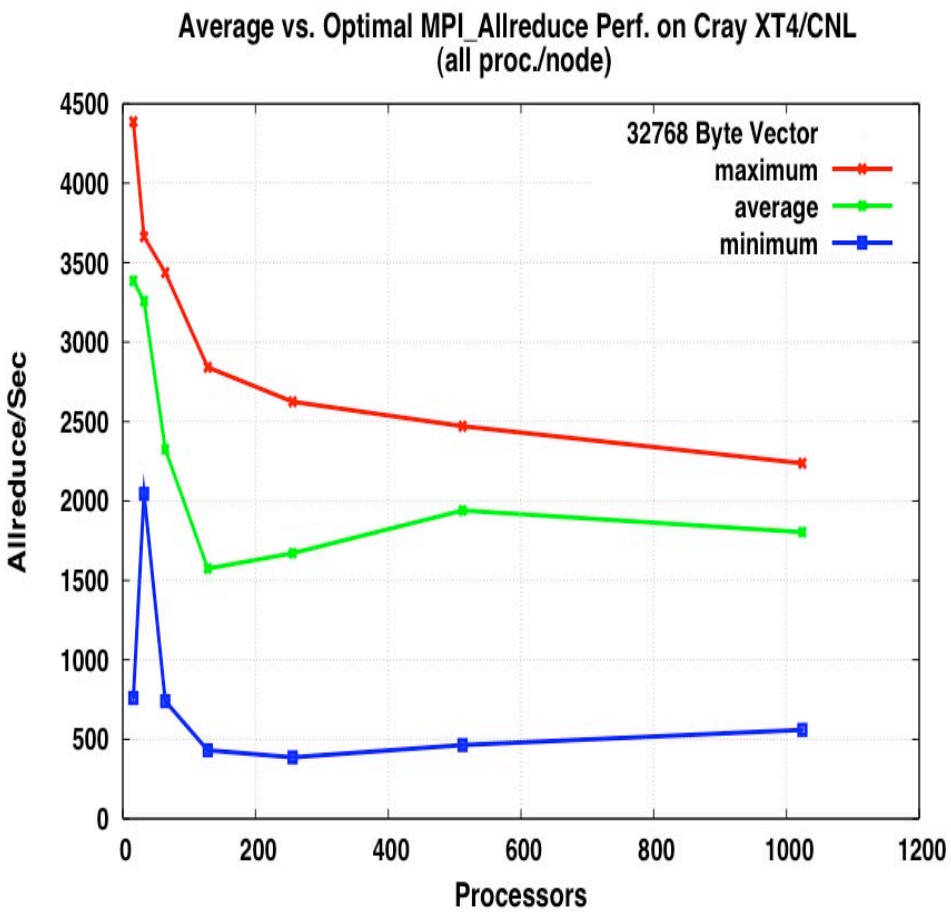
Allreduce test on Franklin (1024B)

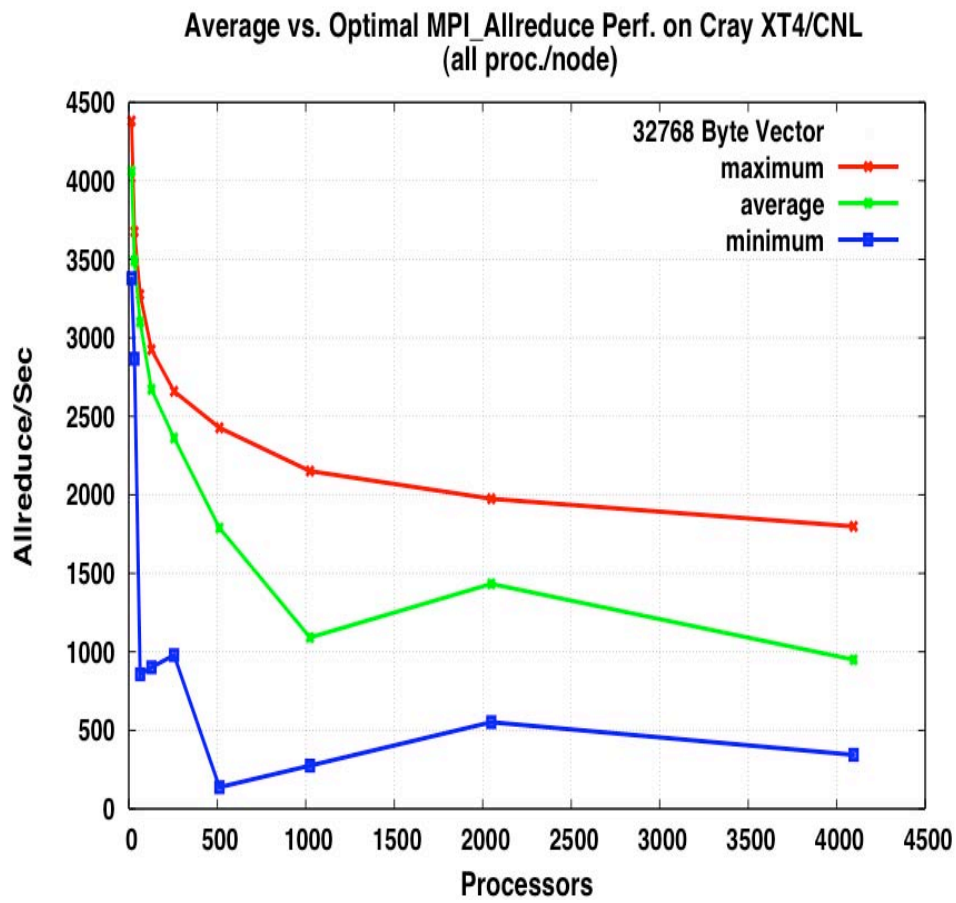




Similar performance issues observed for 1024 Bytes,

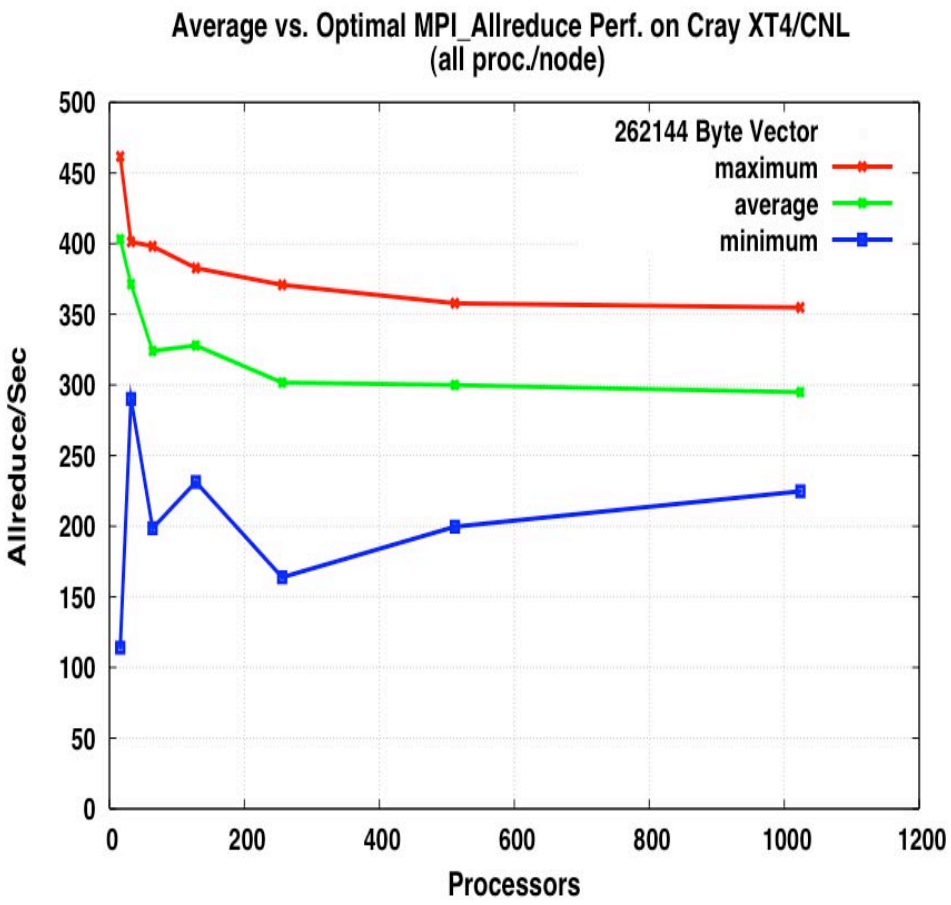
Allreduce test on Franklin (32768B)

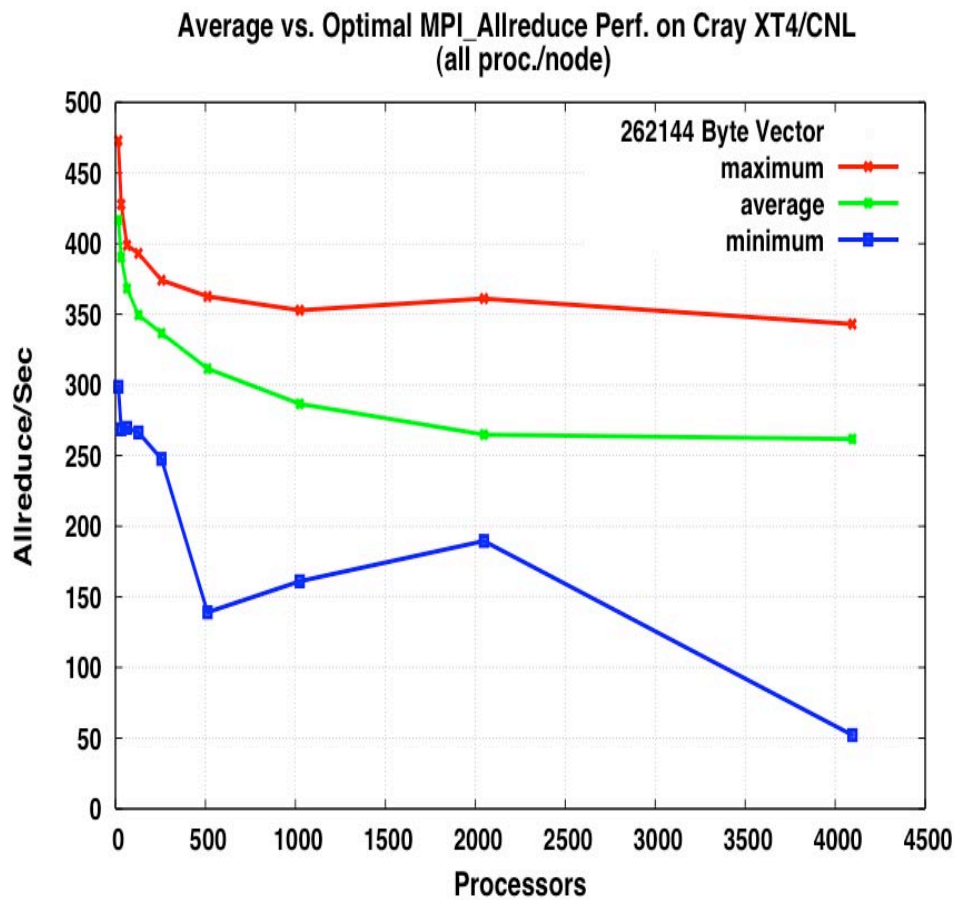




for 32768 Bytes,

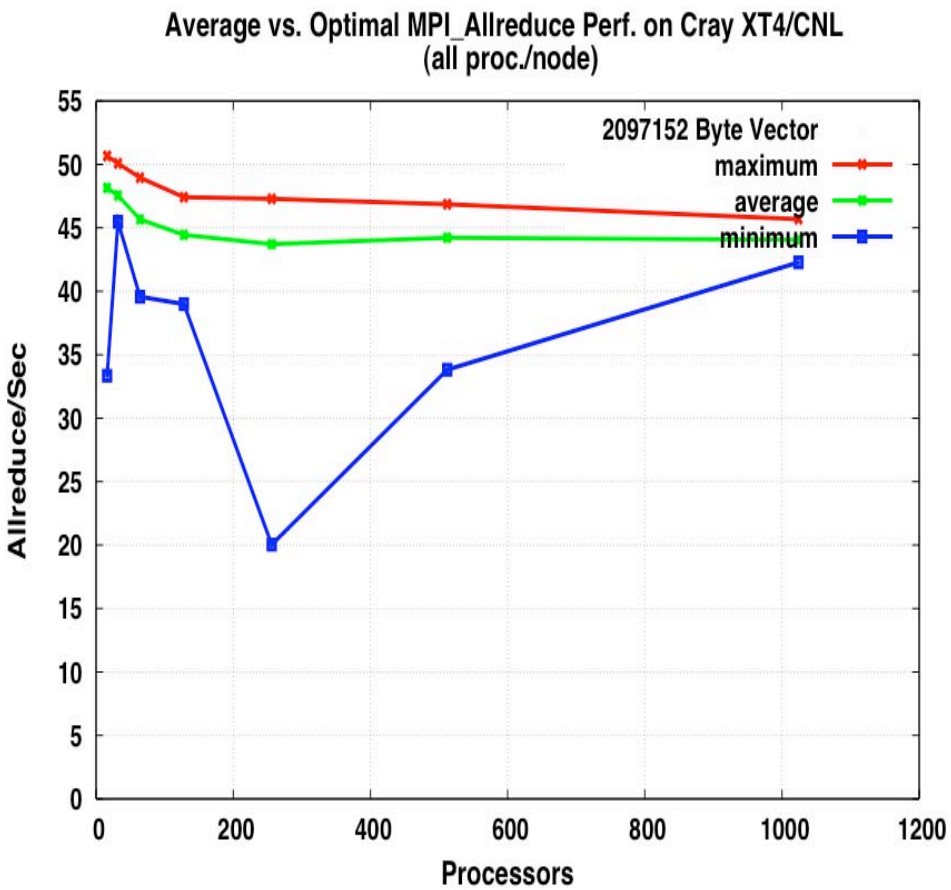
Allreduce test on Franklin (262144B)

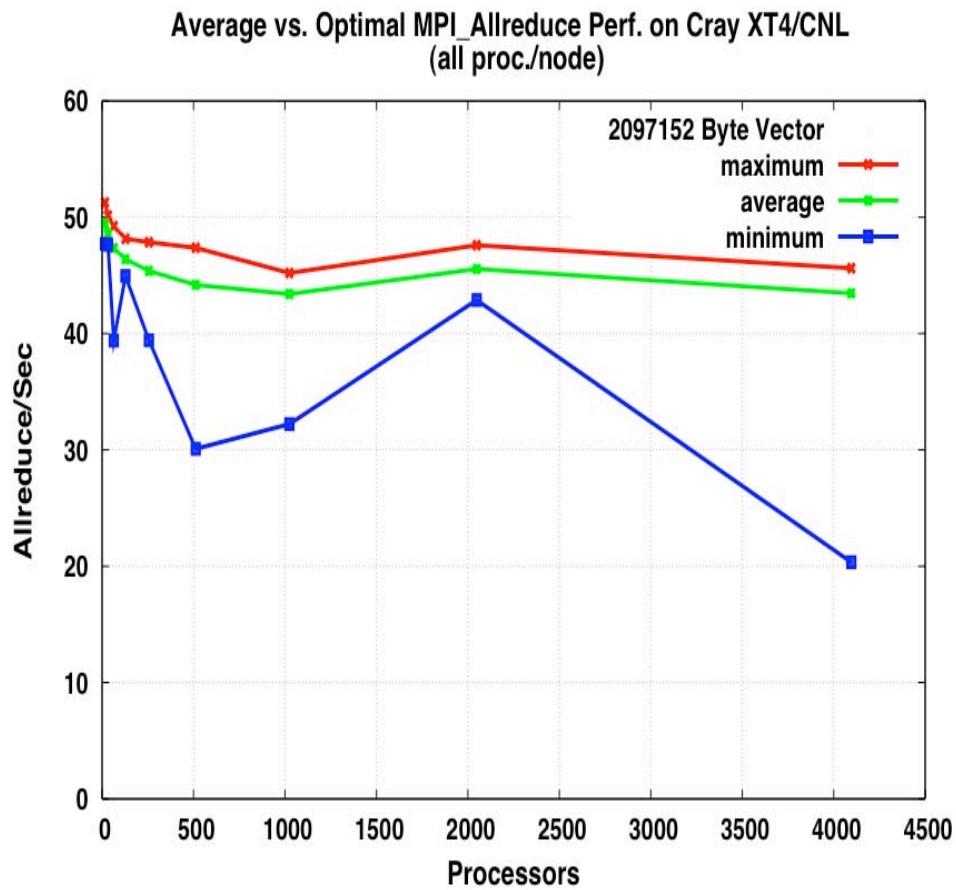




for 262144 Bytes,

Allreduce test on Franklin (2097152B)





and for 2097152 Bytes. Larger vectors are less sensitive (thankfully), but the perturbations do not disappear, so this is NOT the signature that has been typical of jitter in the past. The performance degradation is affecting allreduce regardless of the duration of the event. This appears to be “something else”.

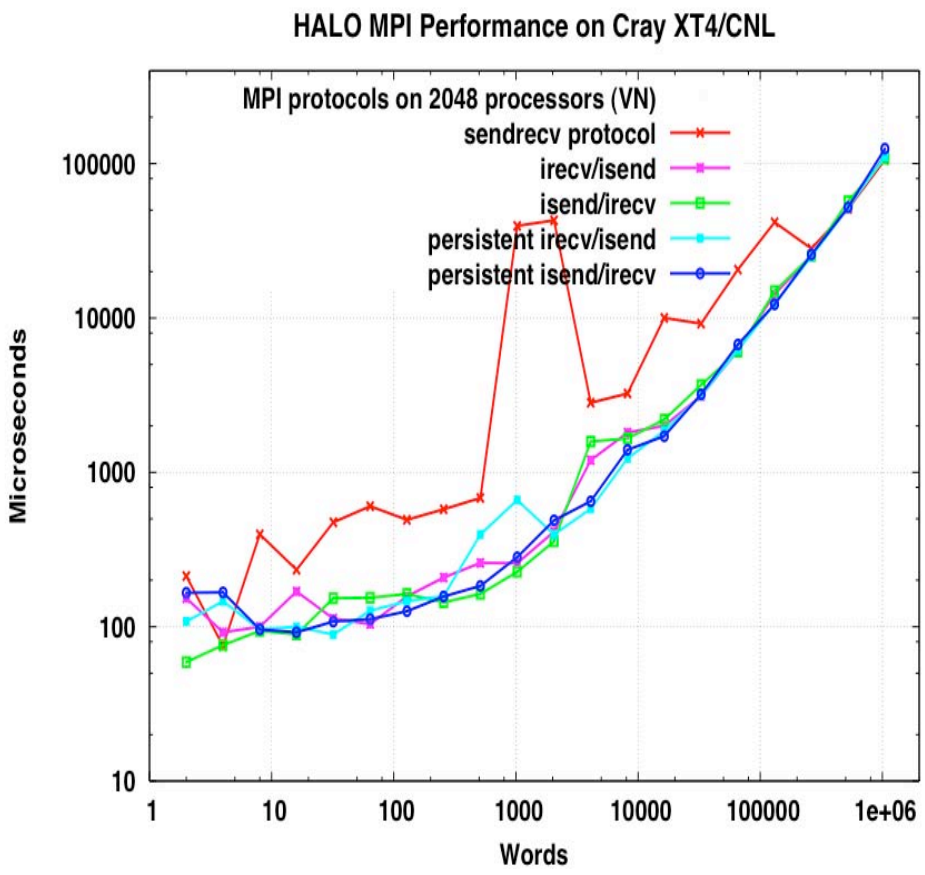
Summary

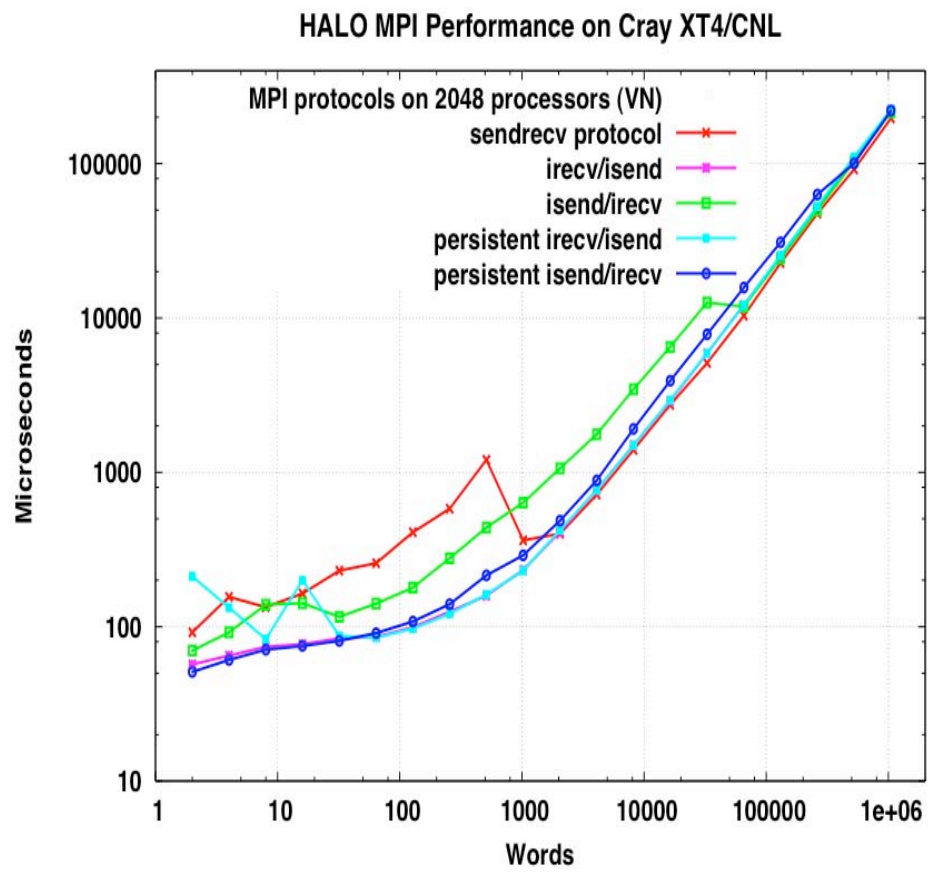
1. Observed significant performance variability on Franklin over these couple of days that was not evident on Jaguarcnl. Does this mean anything?
2. Signature of variability does not appear to be that of OS jitter. Have observed interference from other users when running on ORNL system. Is this what is occurring on Franklin?
3. Doing postmortem analysis, never have enough or the right kind of data. In particular, do not know what else is going on in the system. Need something more/different to more easily diagnose these issues.

Two Months Later

- I have not had time to package up my benchmark codes for NERSC consultants as of yet, but I did just (4/22/08) rerun the HALO and ALLREDUCE experiments. HALO results were less noisy, but cost increased relative to runs in February. Note that the HALO benchmark reports the maximum time (over all processes) after a small number (3-5) of repetitions of the Halo update operator. I generated a modified version that times a single Halo update, and reports max, min, and average …

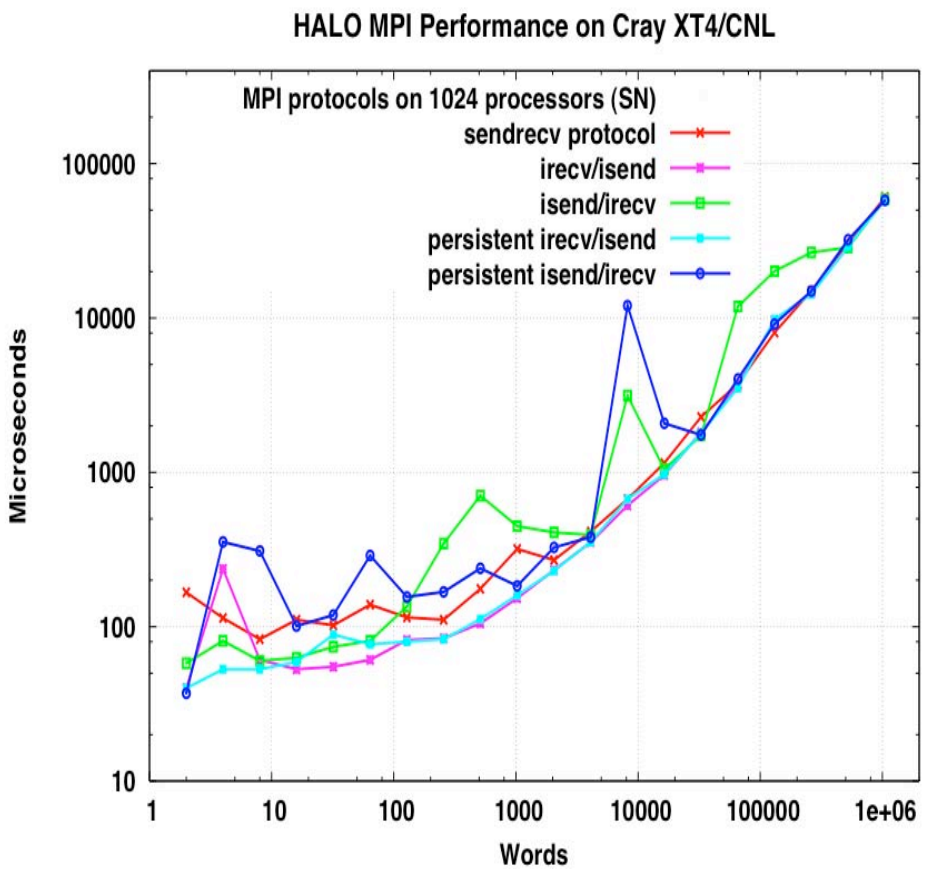
HALO on Franklin: 2/08 vs. 4/08 (2048)

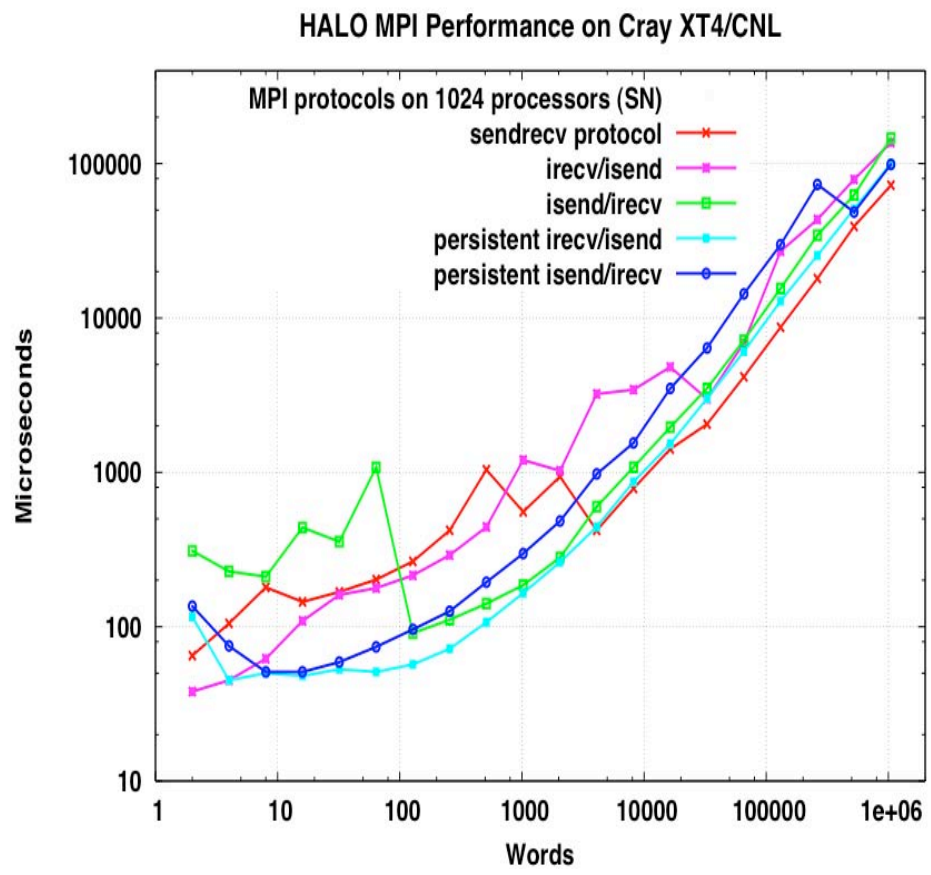




Still some “noise”, but cost for large boundary exchanges is twice what it was before.

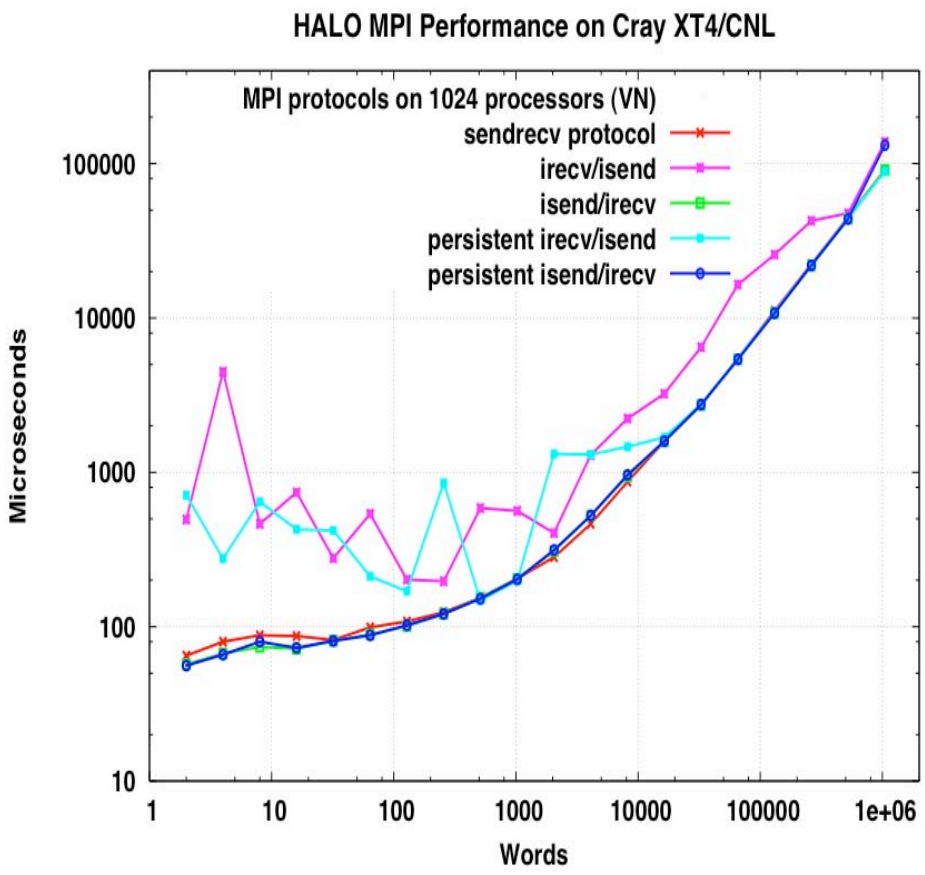
HALO on Franklin: 2/08 vs. 4/08 (1024SN)

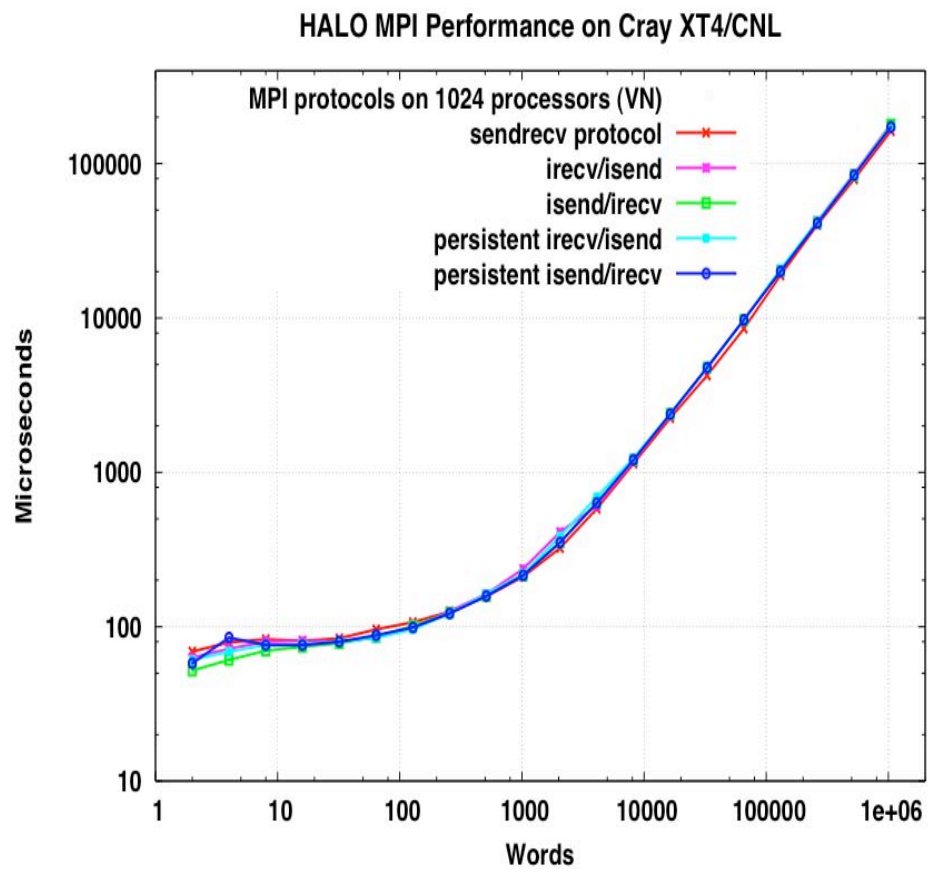




4/08 “1024 SN” experiment occurred in same PBS runs as 2048 VN experiment, so same nodes and near in time.

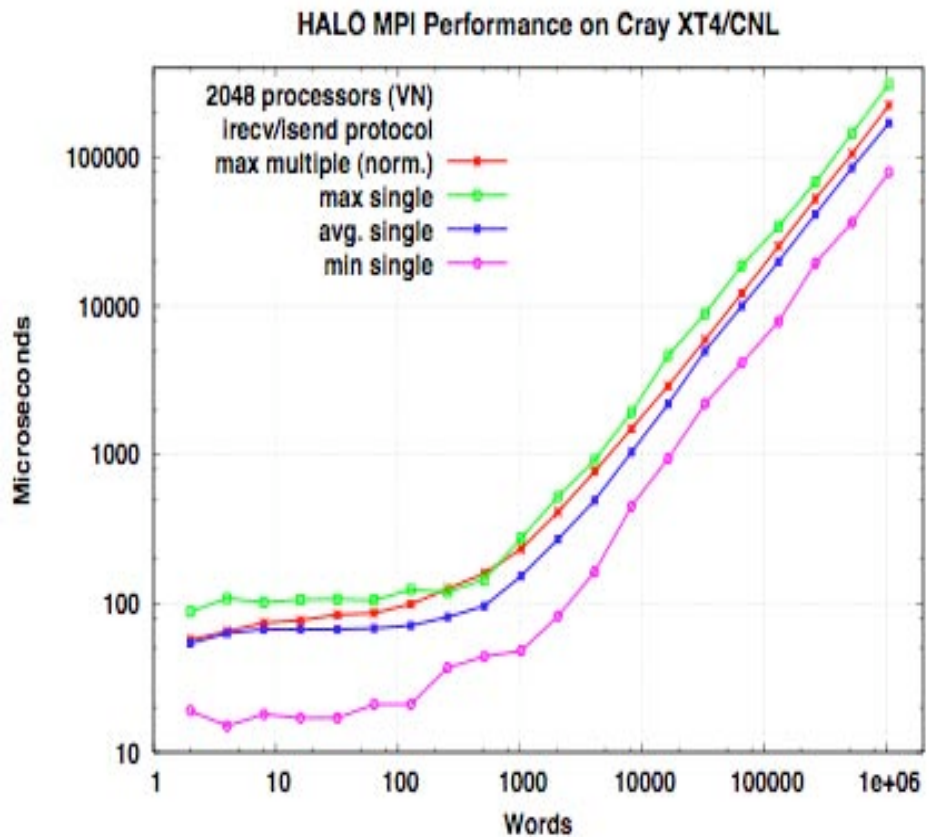
HALO on Franklin: 2/08 vs. 4/08 (4096)

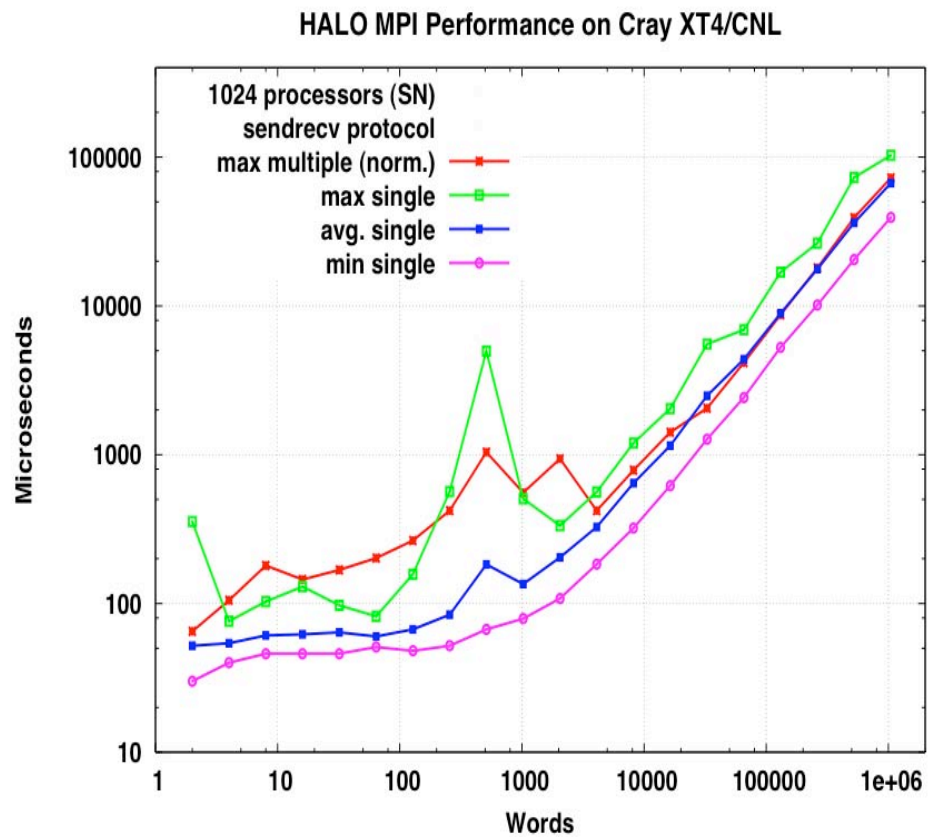




In general, “noise” is less for 4/08 runs, but is also uniformly more expensive. Related? Unrelated? Who knows? How can we determine this?

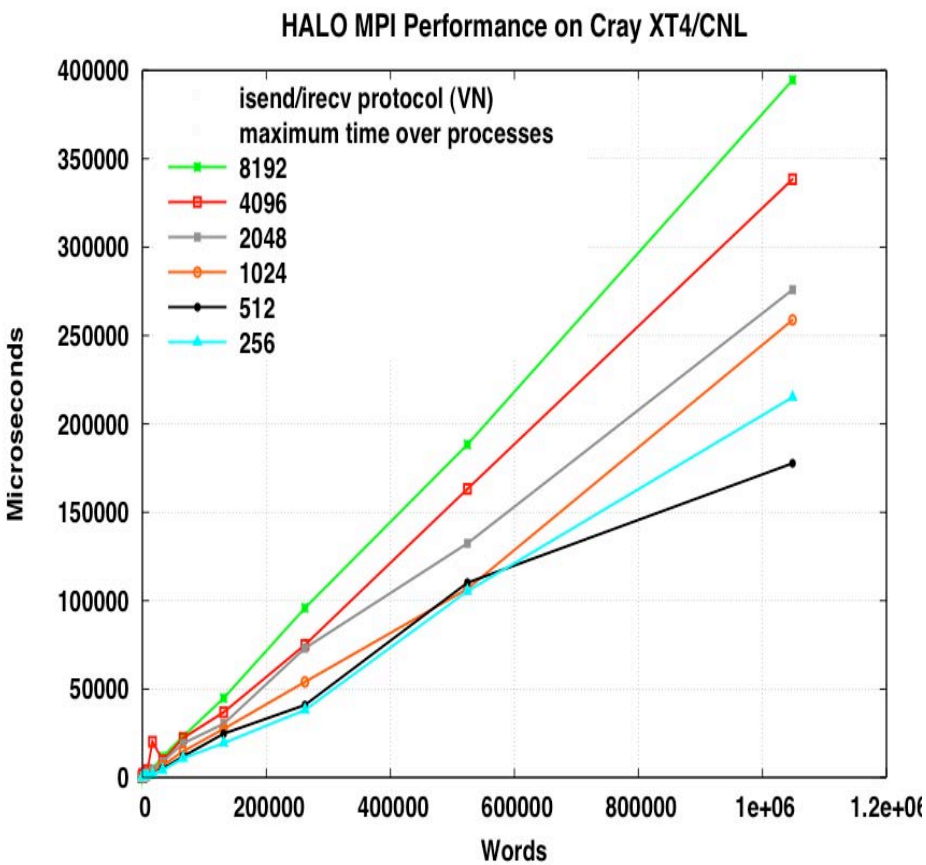
HALO on Franklin: max/avg/min

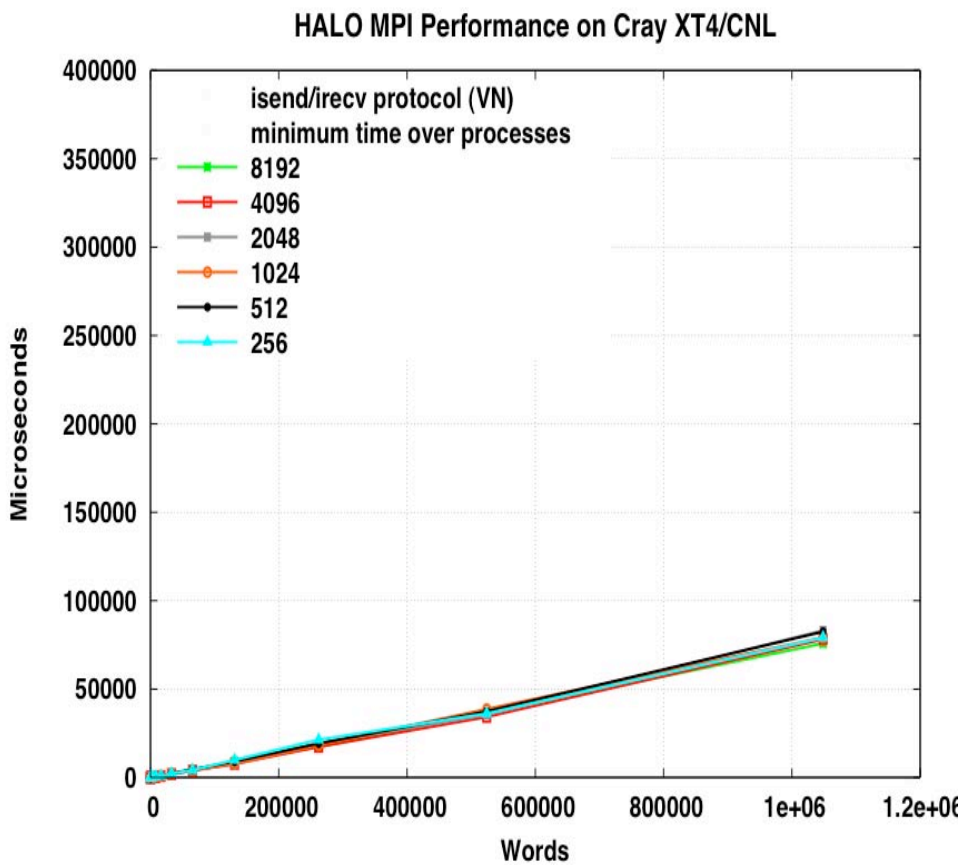




Two examples: (left) no “noise” in either multiple repetition expt or single instance expt., and normalized ‘max’ for multiple rep. is between single instance ‘max’ and single instance ‘avg’. (right) with noise - all in max for single instance

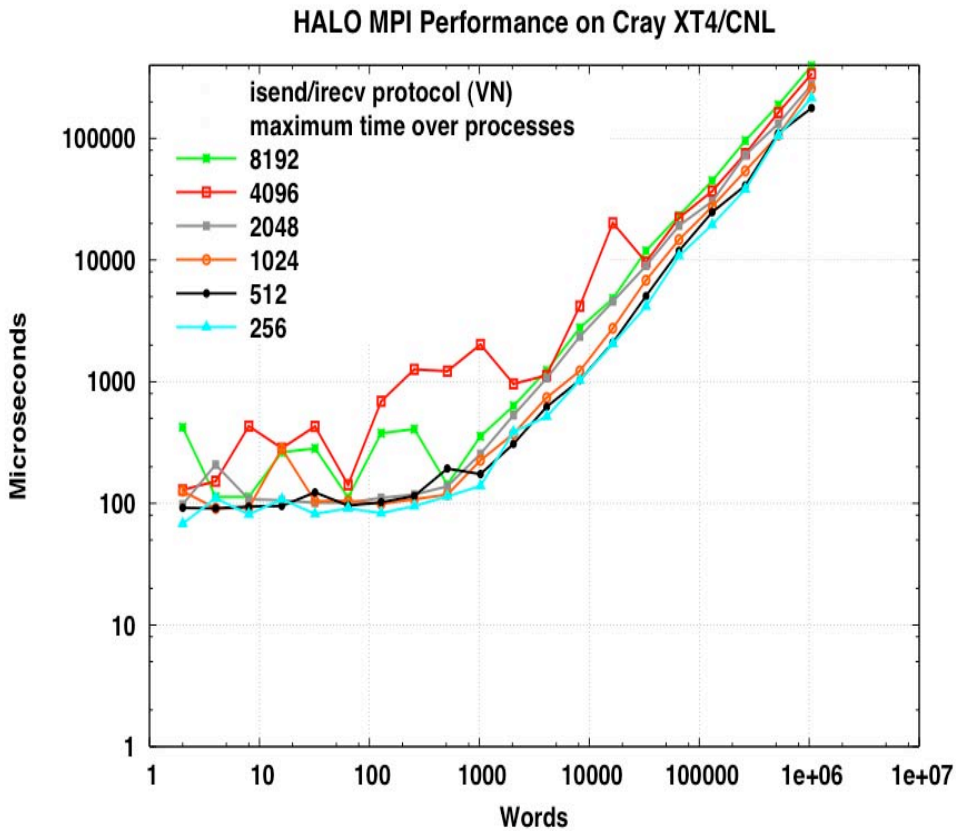
HALO on Franklin: max vs. min

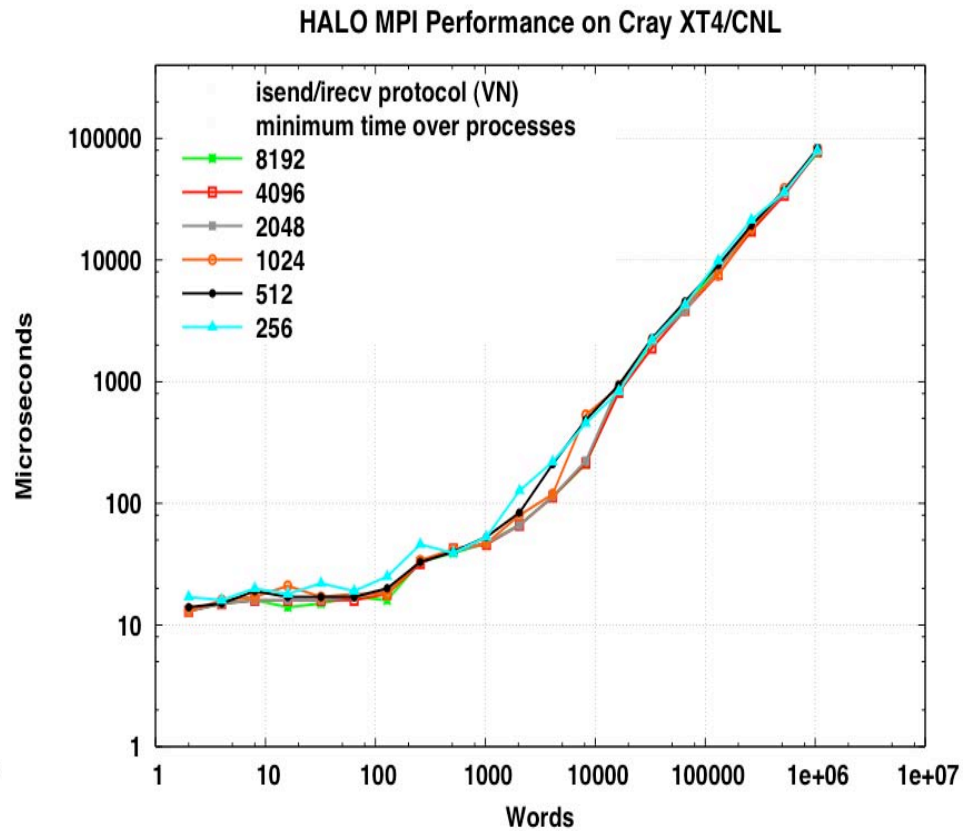




Max. time shows steady growth with total process count (for what should be a local operator). In constrast, minimum time does not show total process count dependence.

HALO on Franklin: max vs. min





Same data as previous slide, but presented in log-log plots. Max. time graph also generally shows increased “noise” with process count, and time for small halos is a factor of 5-10 greater than that of the minimum time.

Summary - 4/24/08

- What is going on? OS jitter? Slow nodes? Interference from neighbors (who are contending for link bandwidth)? How can I tell? Nature of variability from run to run implies that not *just* OS jitter? Consistent gap (growing as a function on process count) between min and max indicates that OS jitter-type issues do play a role? Would like to be able to differentiate between the possible issues. I need help (global system information) to be able to do this, either that or dedicated time.