AutoNUMA

Red Hat, Inc.

Andrea Arcangeli
aarcange at redhat.com

30 Aug 2012

redhat.

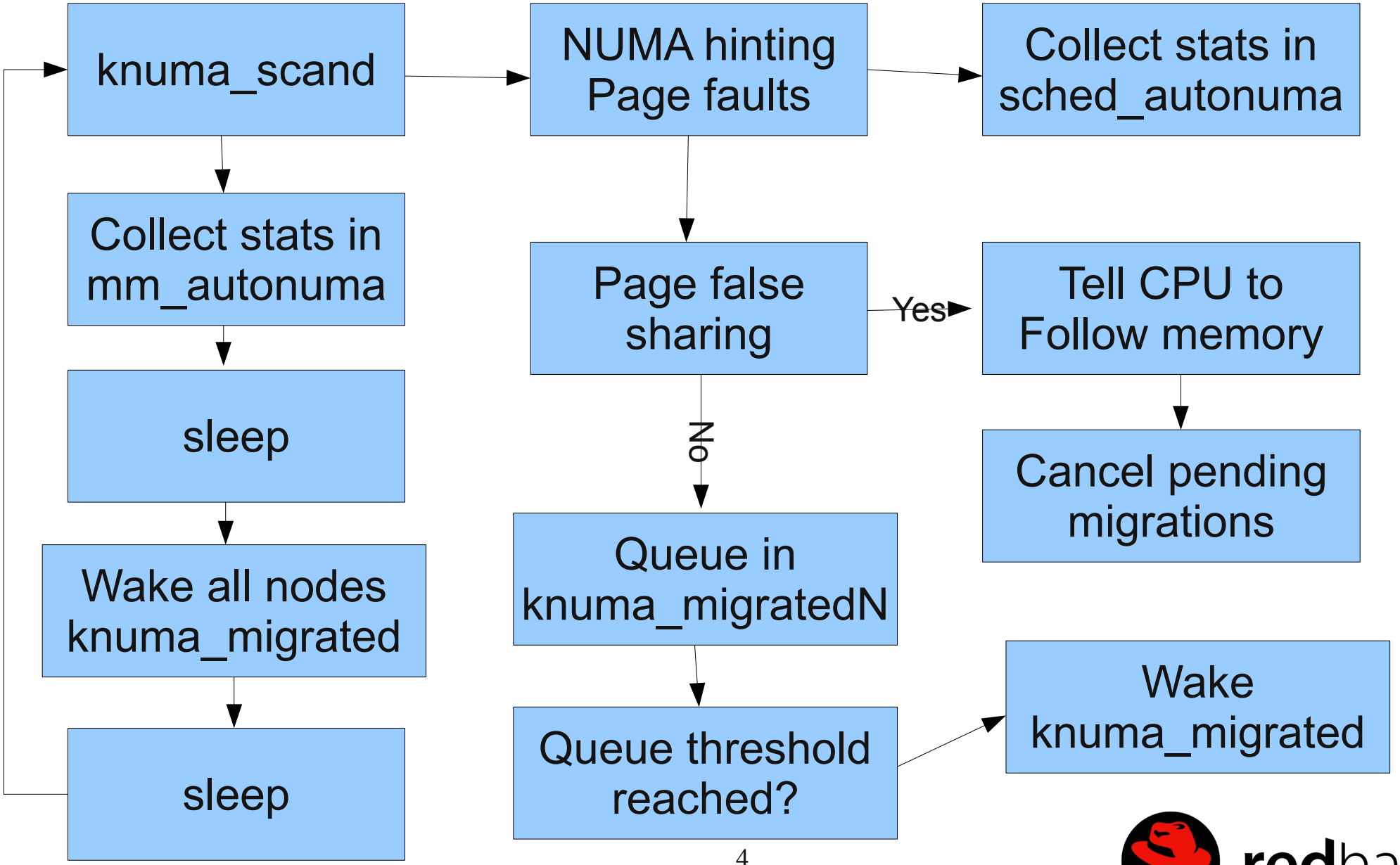
AutoNUMA components

- knuma_scand
  - If stopped, everything stops
  - Triggers the chain reaction when started
- NUMA hinting page faults
- knuma_migratedN (per node)
- scheduler (CPU follow memory & active idle balancing)
- Memory follow CPU (NUMA hinting page faults)
- False sharing detection

AutoNUMA data

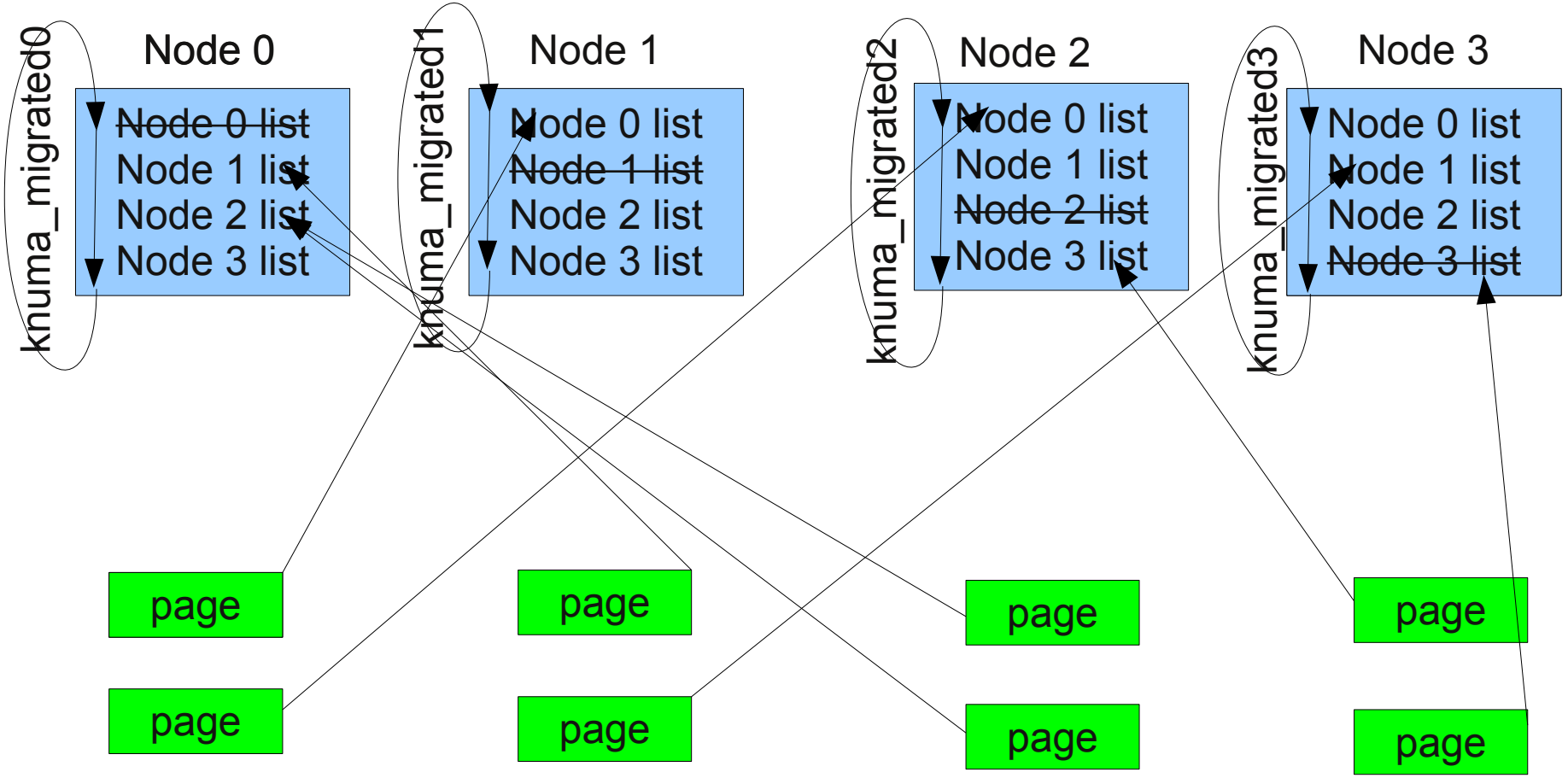
- sched_autonuma
  - task_struct
- mm_autonuma
  - mm_struct
    - Working set or ~RSS

AutoNUMA logic



redhat

AutoNUMA knuma_migratedN

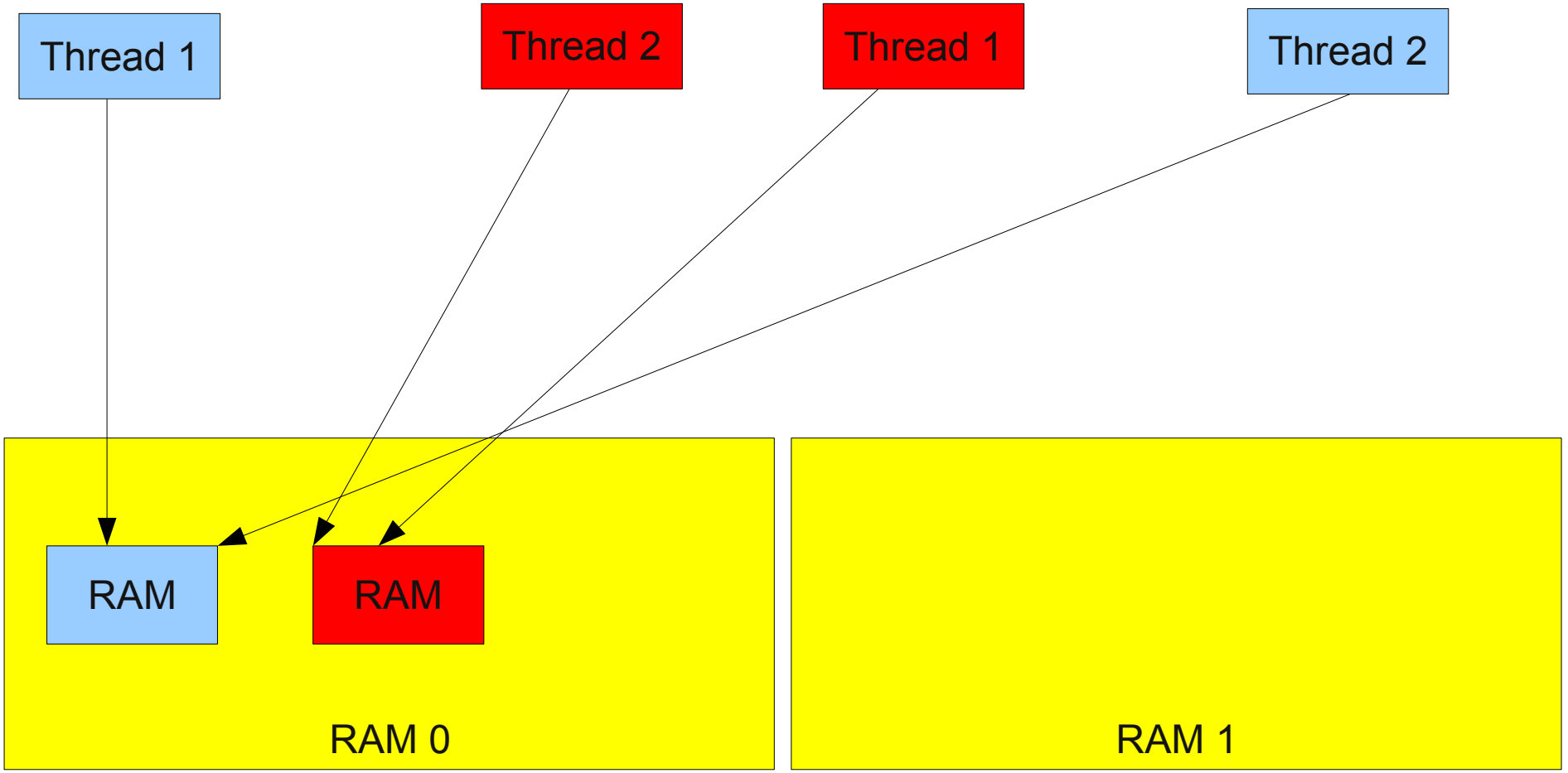


redhat.

AutoNUMA-benchmark tests

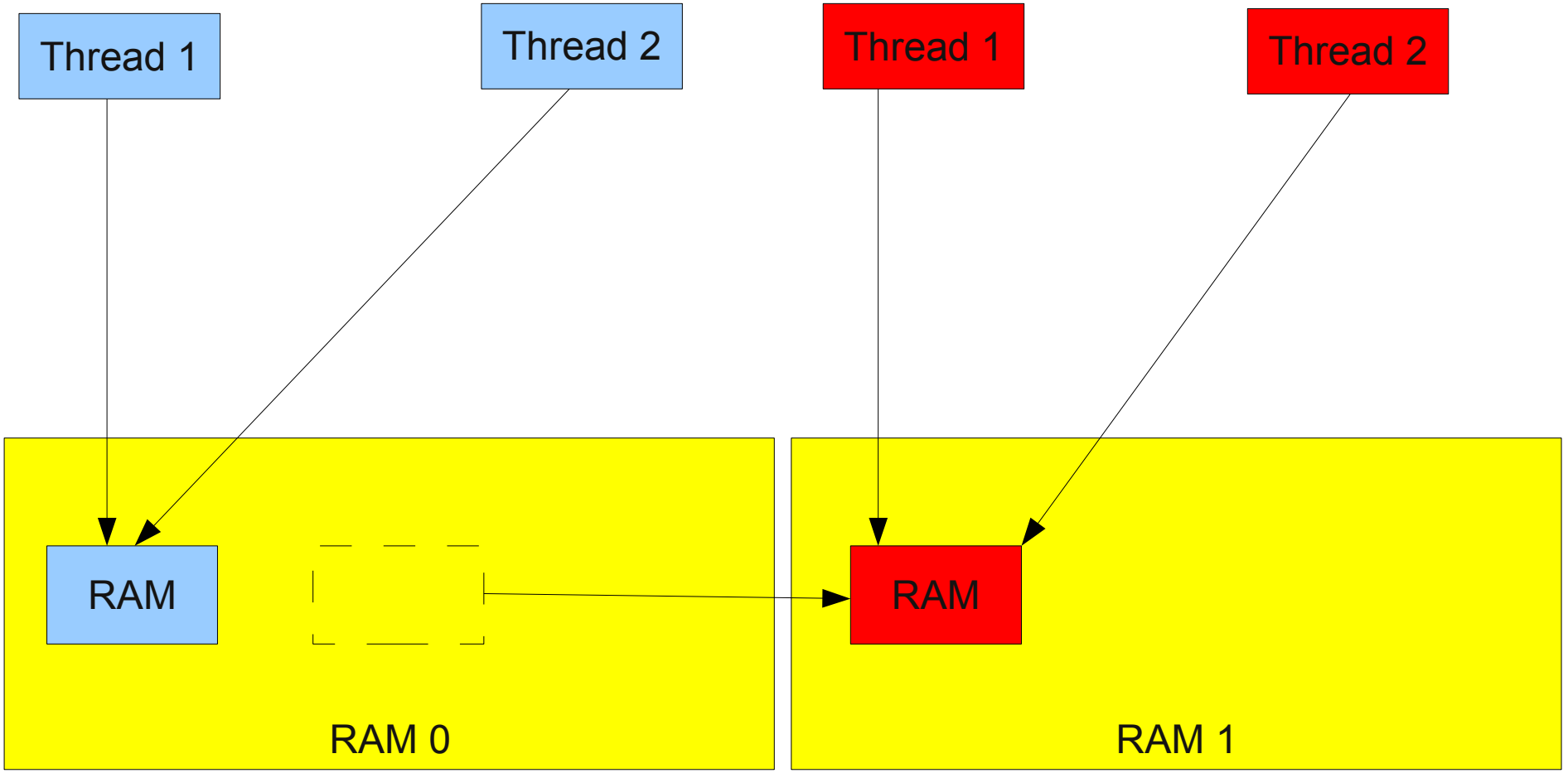
- P=# of processes, T=# of threads per process, M=memory per process
- numa01
  - M=3GiB P=2, T=nr_cpus/2, all threads share all process memory
- numa01_THREAD_ALLOC
  - M=3GiB, P=2, T=nr_cpus/2, all threads use per-thread local memory
- numa02
  - M=1GiB, P=1, T=nr_cpus, all threads use per-thread local memory
- numa02_SMT
  - M=1GiB, P=1, T=nr_cpus/2, all threads use per-thread local memory
    - The kernel to get this right must not use more than one HT thread per core, and in turn it must decide to split the load over two NUMA nodes even if the load would fit in a single NUMA node
    - Testing with T=nr_cpus/4 and smaller T values, would also be interesting, but if the kernel behaves well with T=nr_cpus/2 there's a good chance it'll behave sanely with more than half of the CPUs idle too
- More will be added...

numa01 (startup)



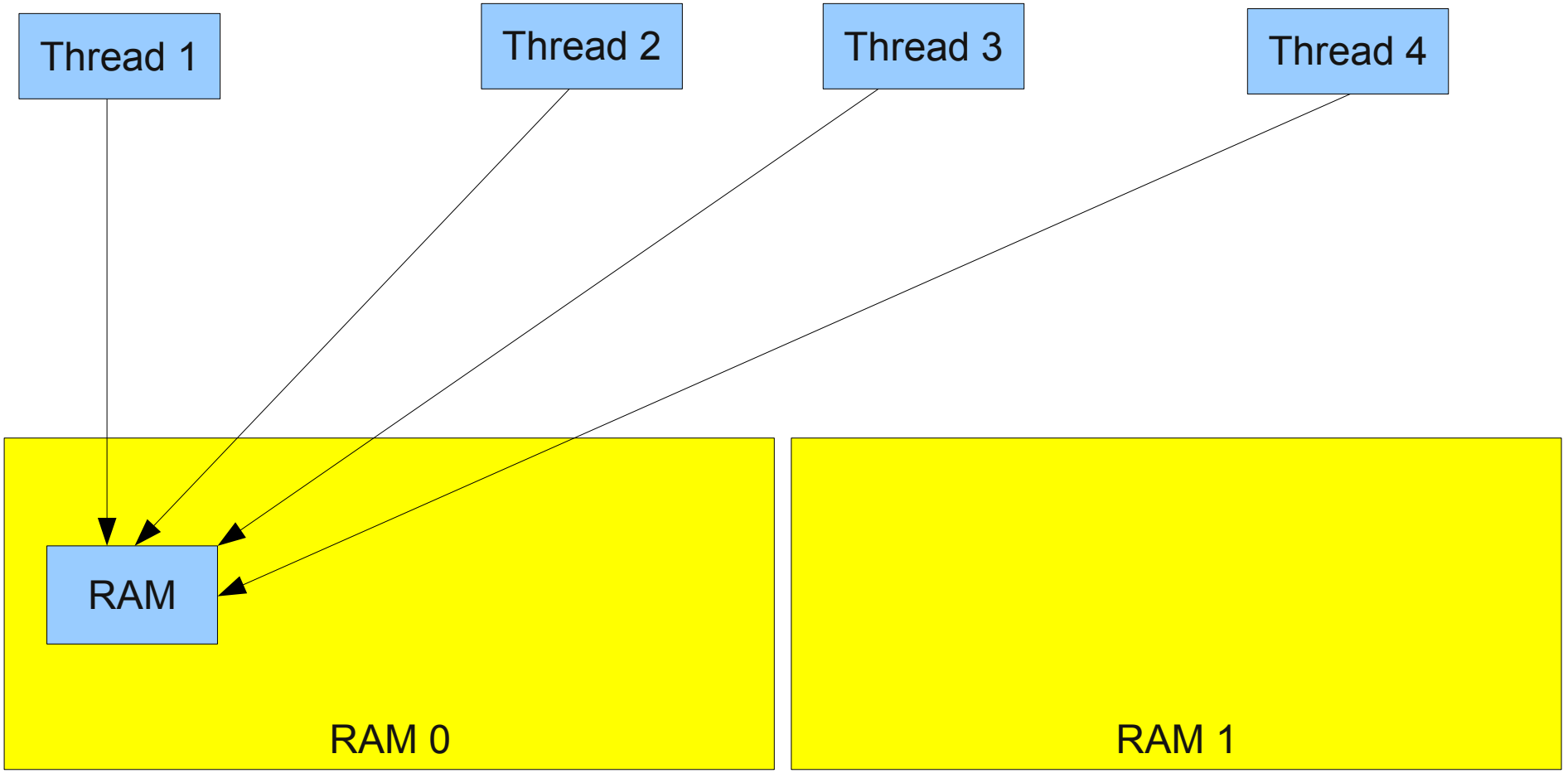
redhat

numa01 (converged)



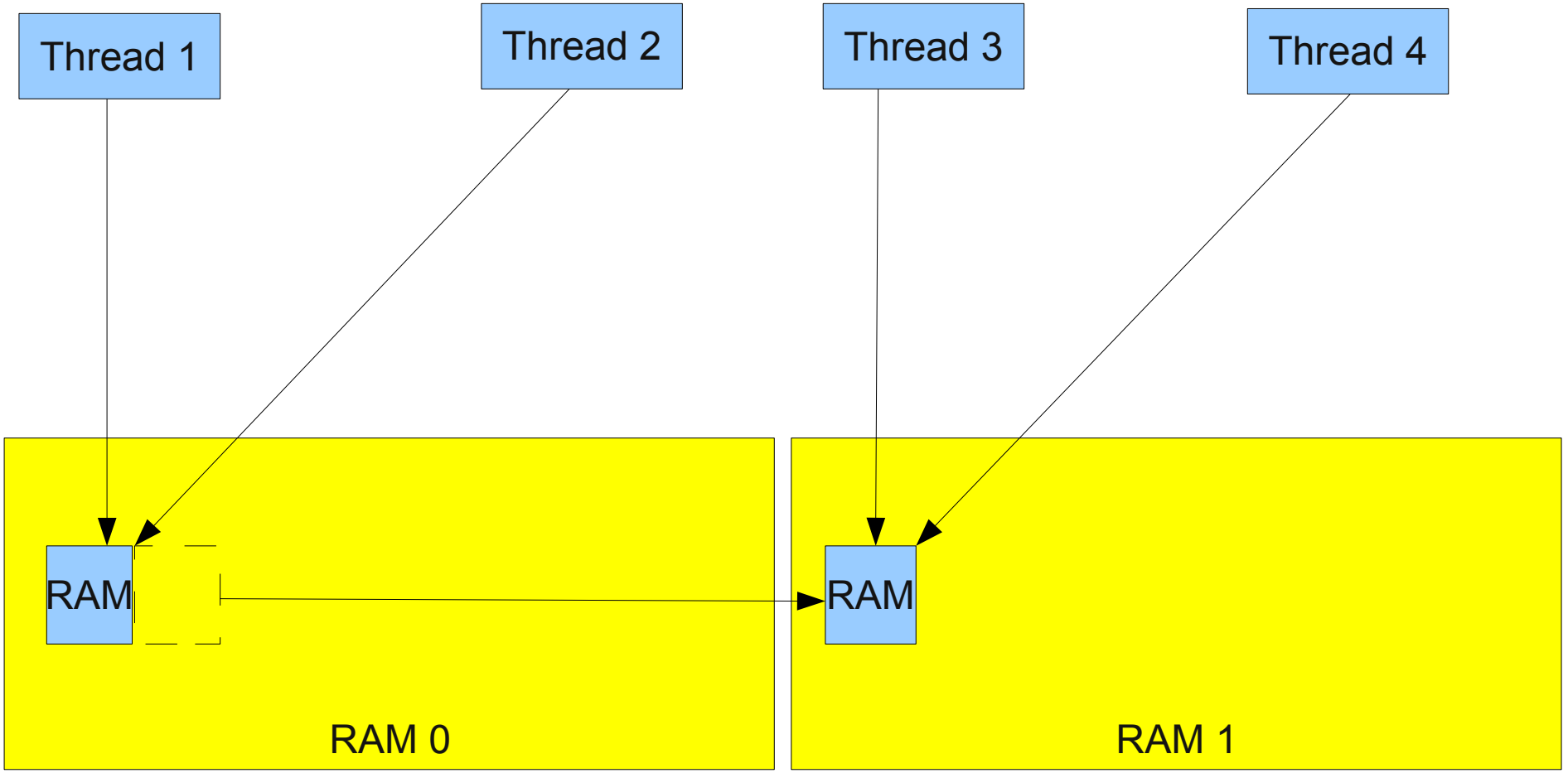
redhat

numa02 (startup)



redhat

numa02 (converged)



redhat

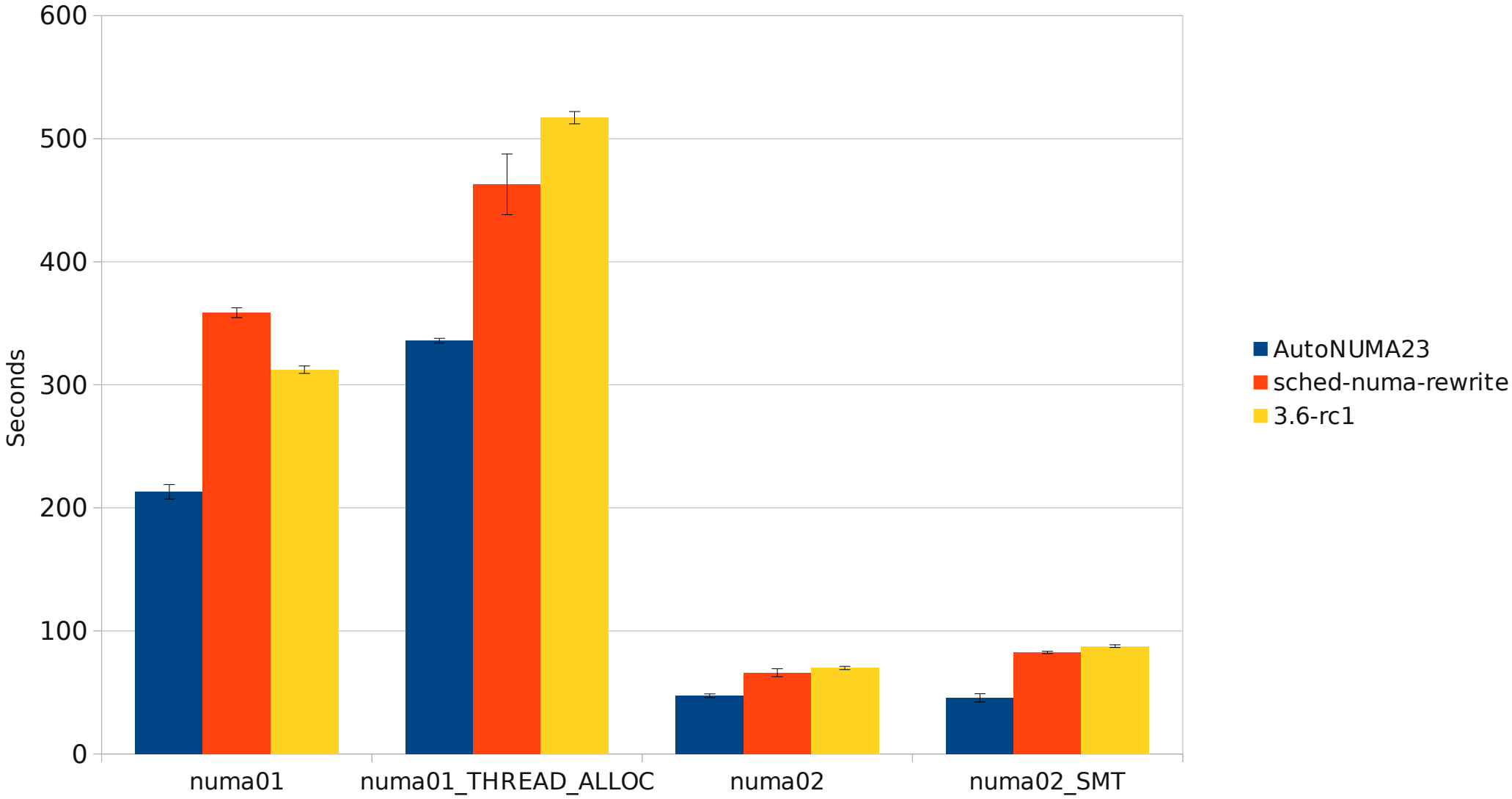
autonuma-benchmark

```
$ git clone git://gitorious.org/autonuma-benchmark/autonuma-benchmark.git
$ cd autonuma-benchmark
$ sudo ./start_bench.sh -s -t
```


redhat.

4 runs, average and stdev

Lower is better

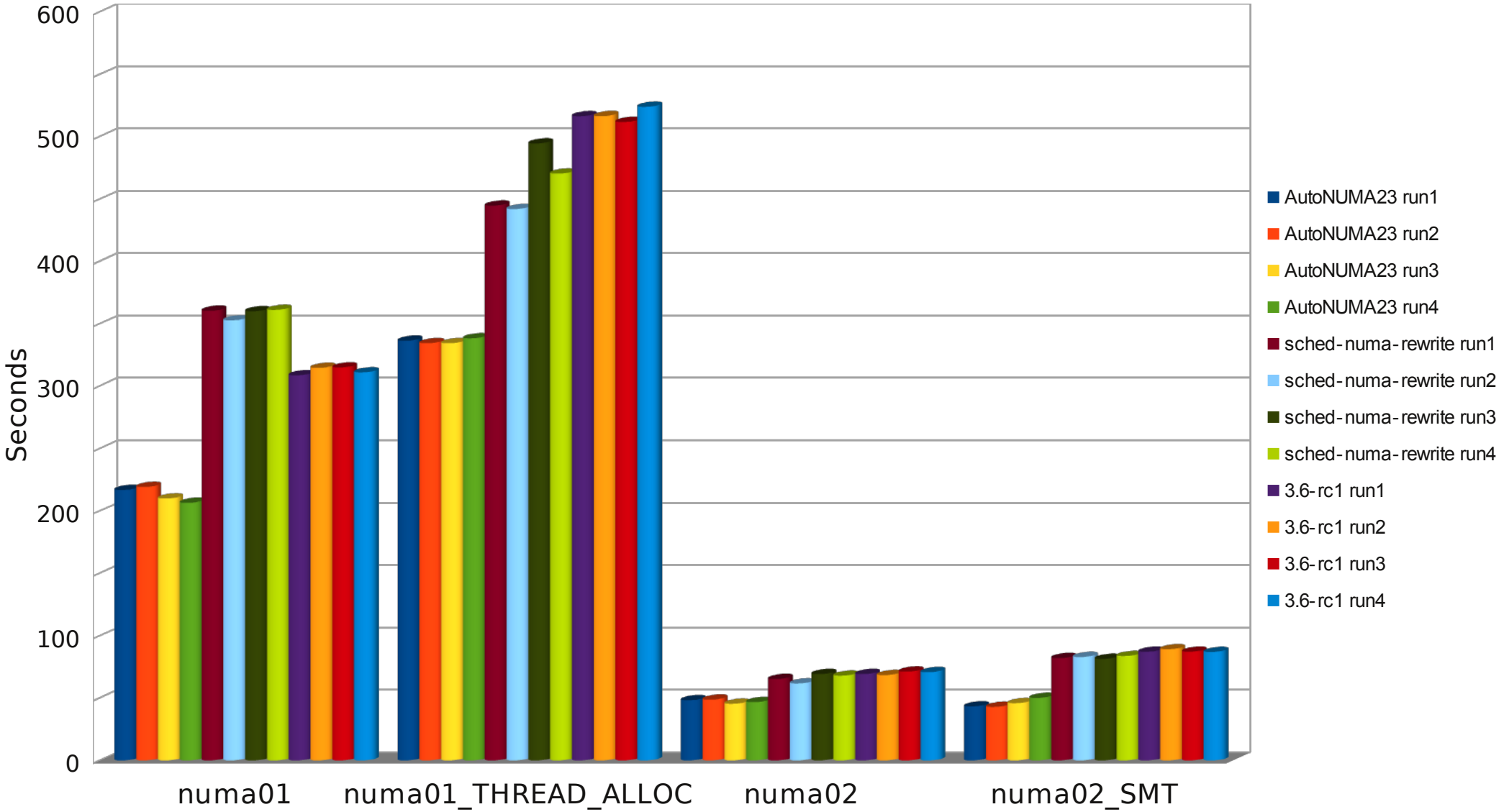


2 nodes

redhat.

4 runs of each test, all results

Lower is better

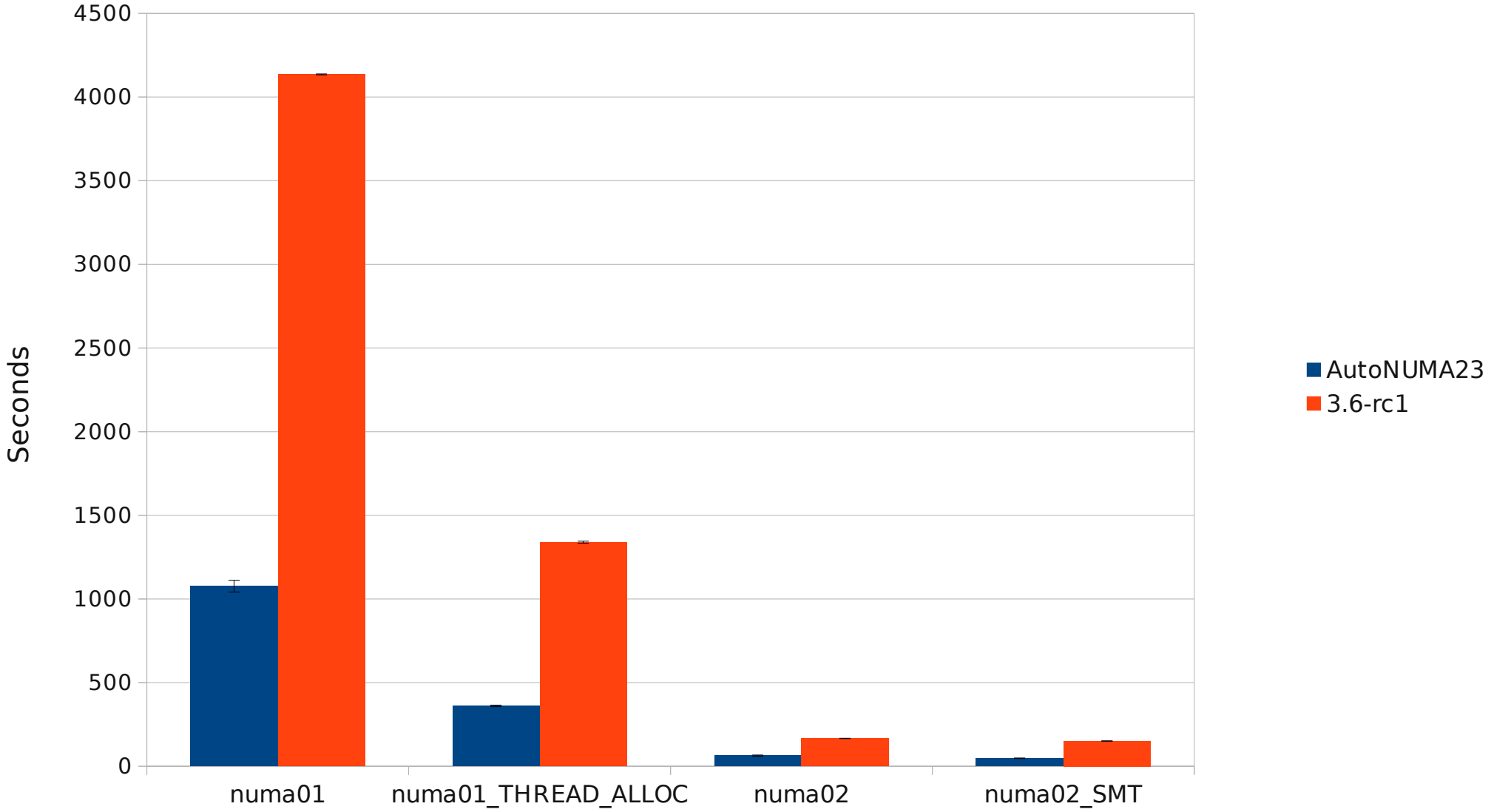


2 nodes

redhat.

4 runs, average and stdev

Lower is better

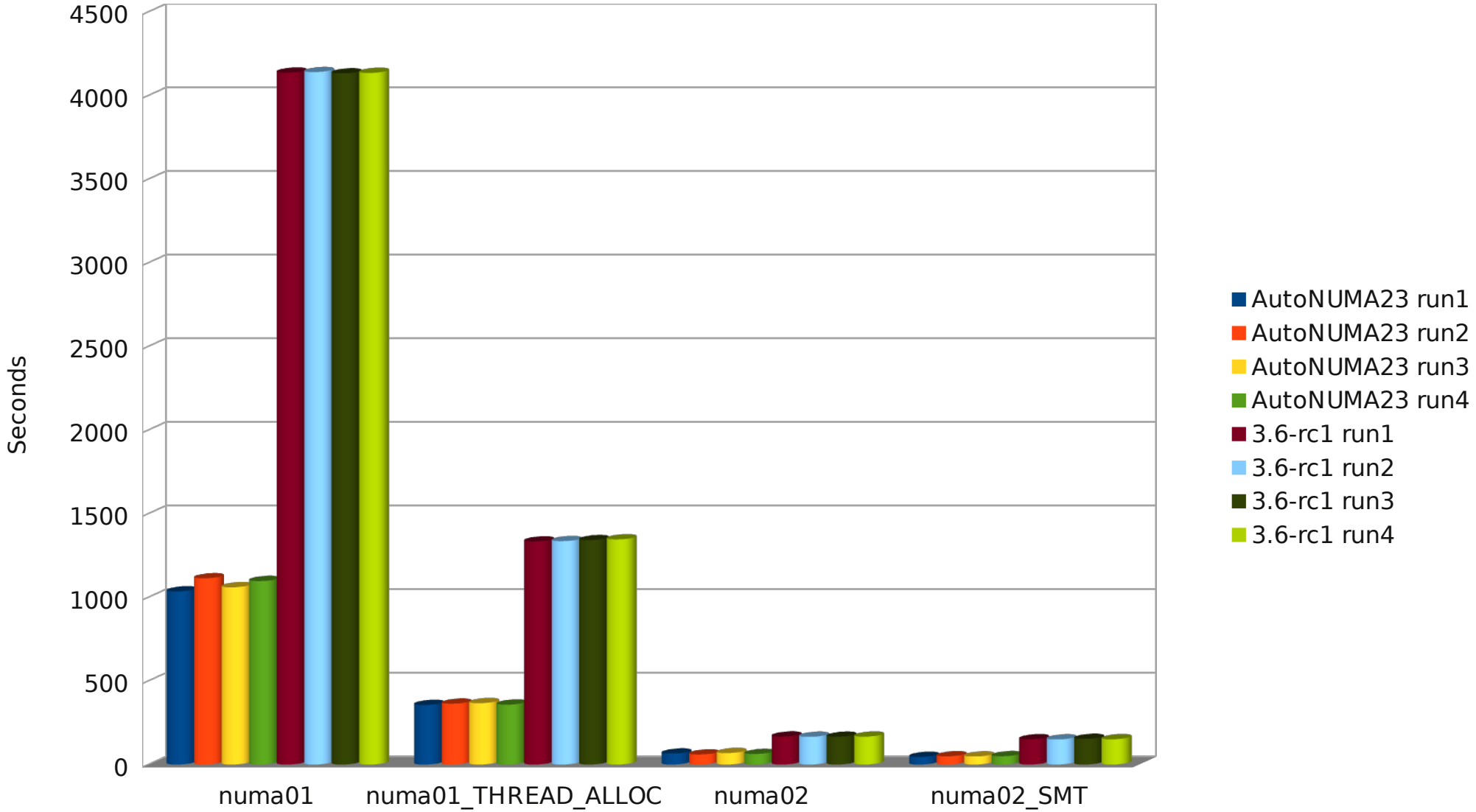


8 nodes

redhat.

4 runs of each test, all results

Lower is better



8 nodes

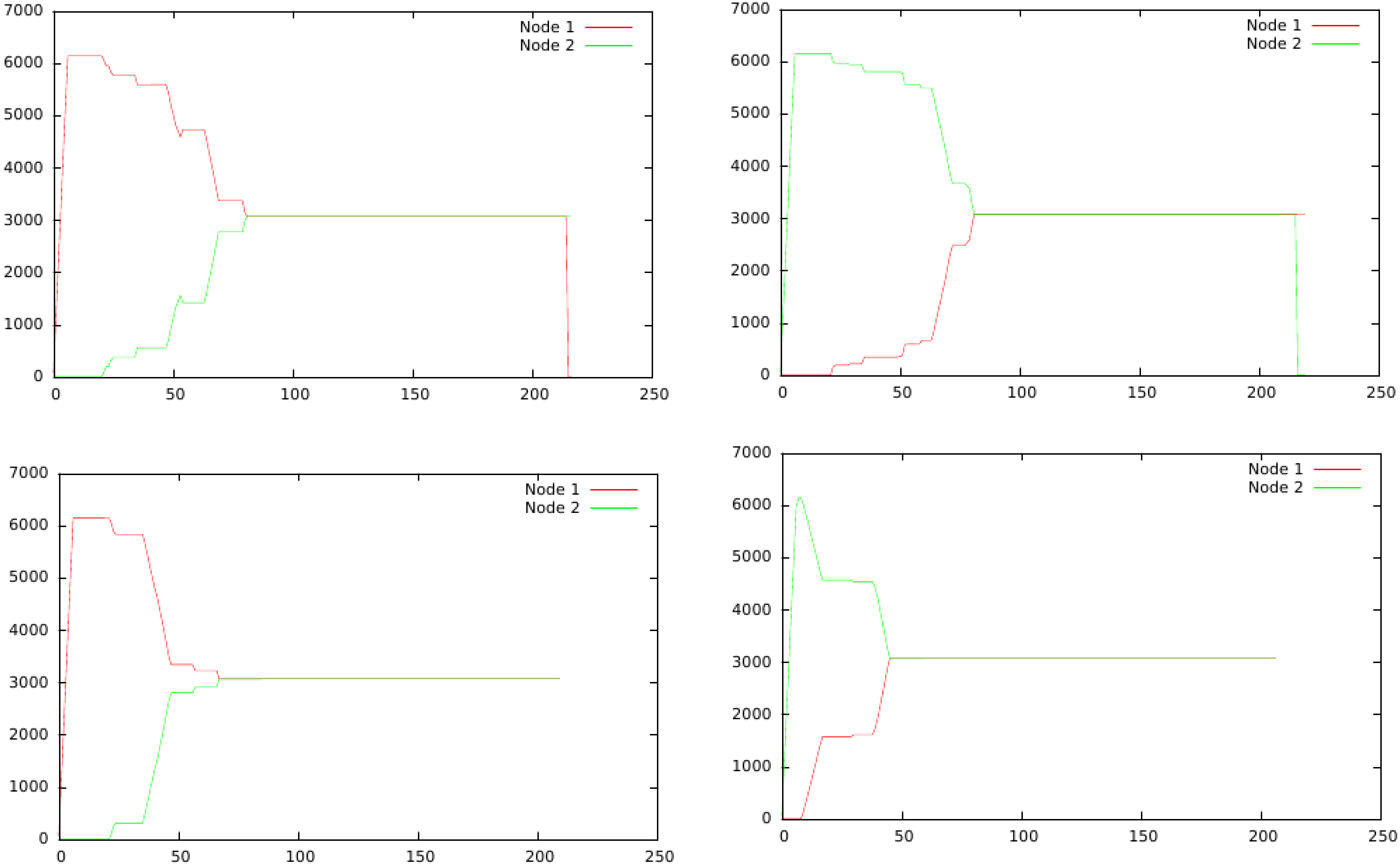
redhat.

Convergence charts

- The AutoNUMA-benchmark produces a chart for each test run (in pdf format).
  - In each chart there is one two dimensional line per NUMA node (node1, node2, etc.).
    - X=time (seconds)
    - Y=memory (MiB)
  - Each line represents how much of the test process's memory is in that NUMA node over time, for the duration of the test. Because all memory starts in one node, it illustrates how the memory migrates over time.
  - A workload converges when the memory levels are equal in all NUMA nodes
  - Note: with nr_nodes > 2, numa01 may not fully converge because half of the cpus will thrash on the memory of half of the nodes, but it should get close enough
  - In the future we plan to add a new numa01 "PER_NODE" test with P=nr_nodes and T=nr_cpus/nr_nodes

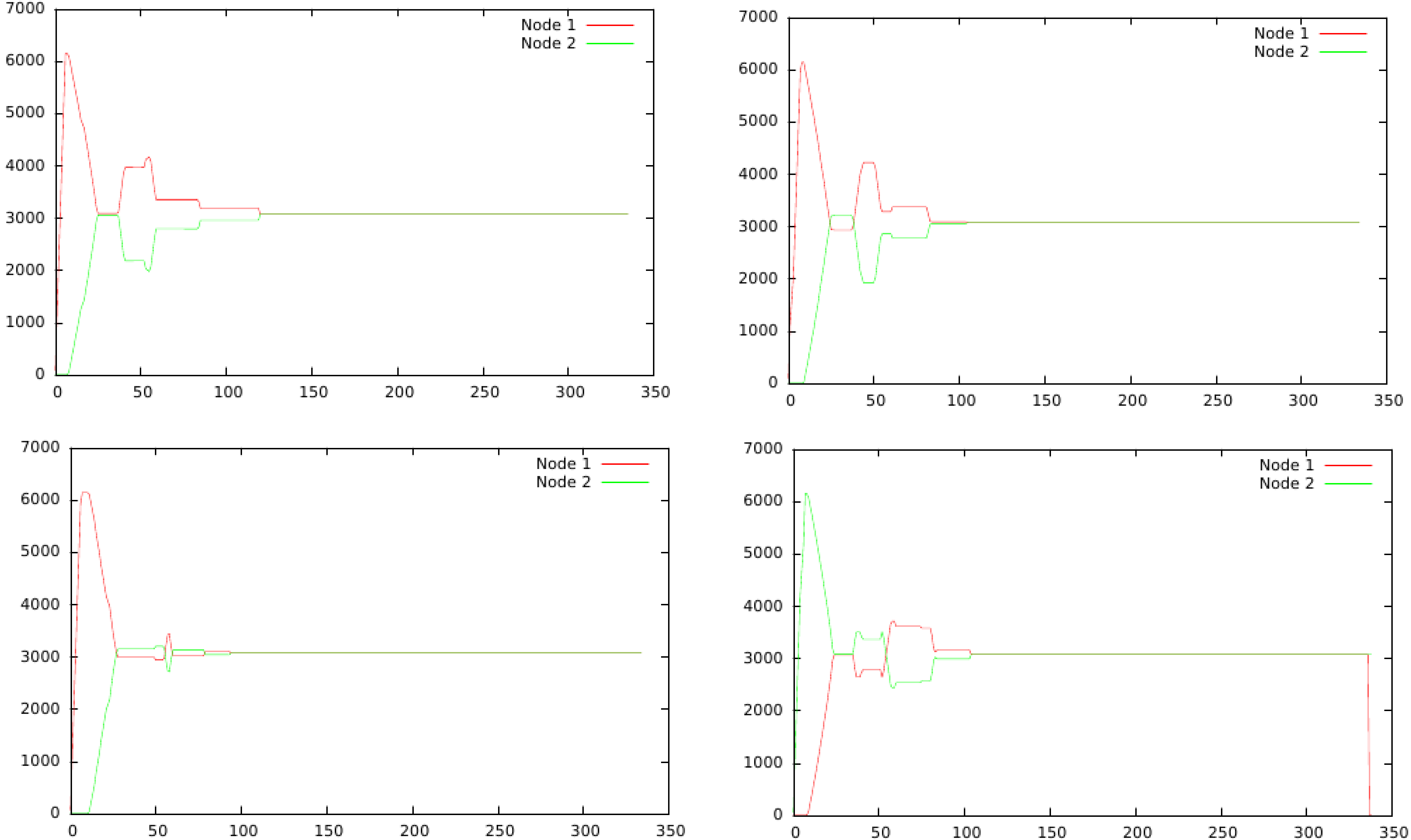
redhat.

AutoNUMA23 numa01



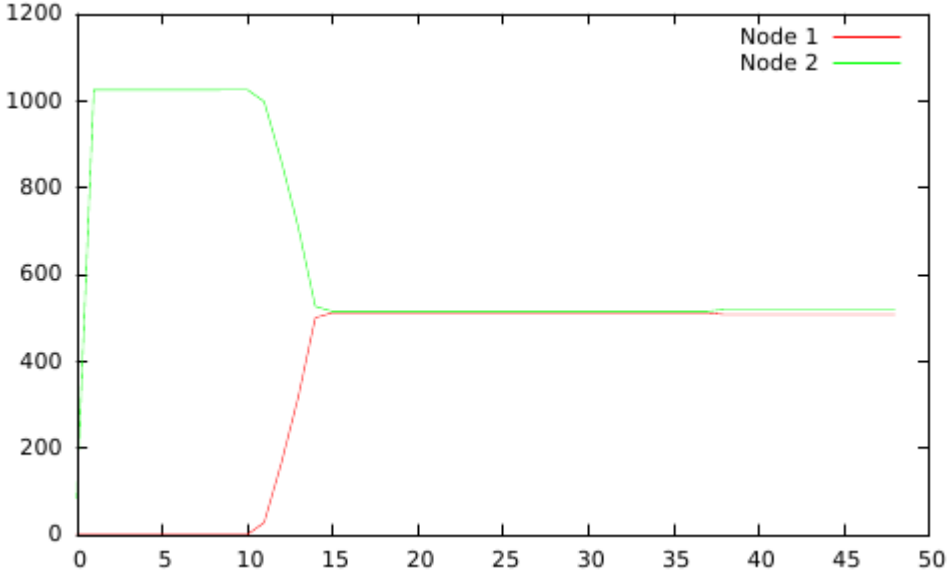
2 nodes

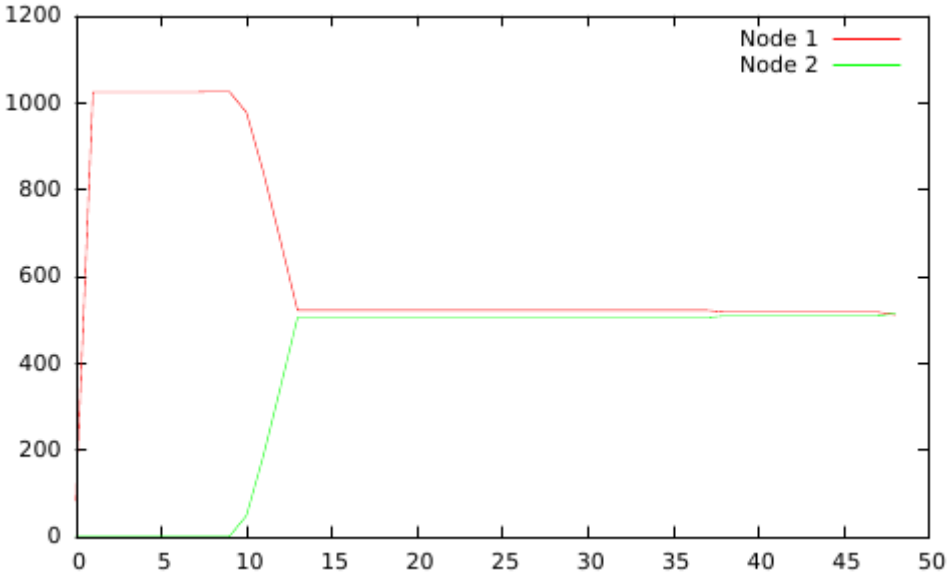
AutoNUMA23 numa01_THREAD..

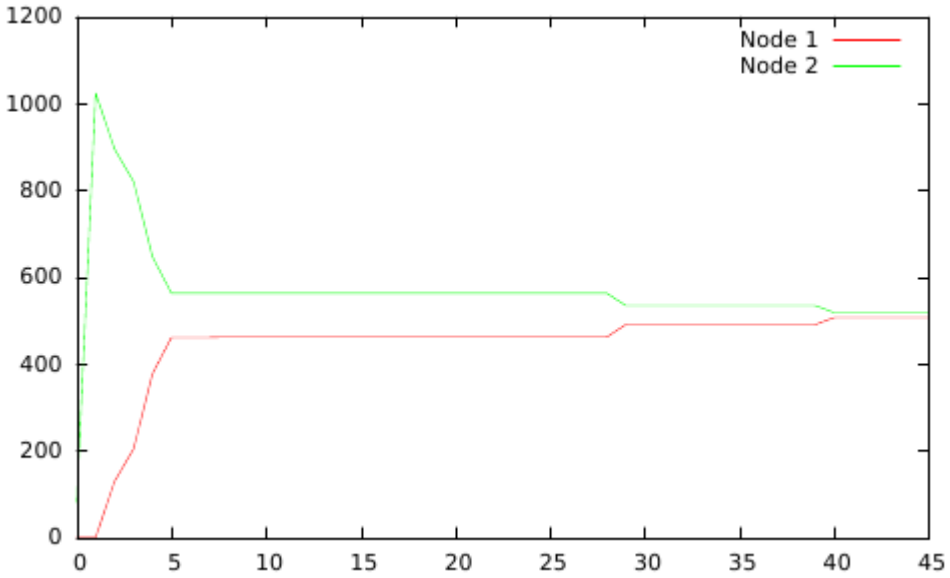


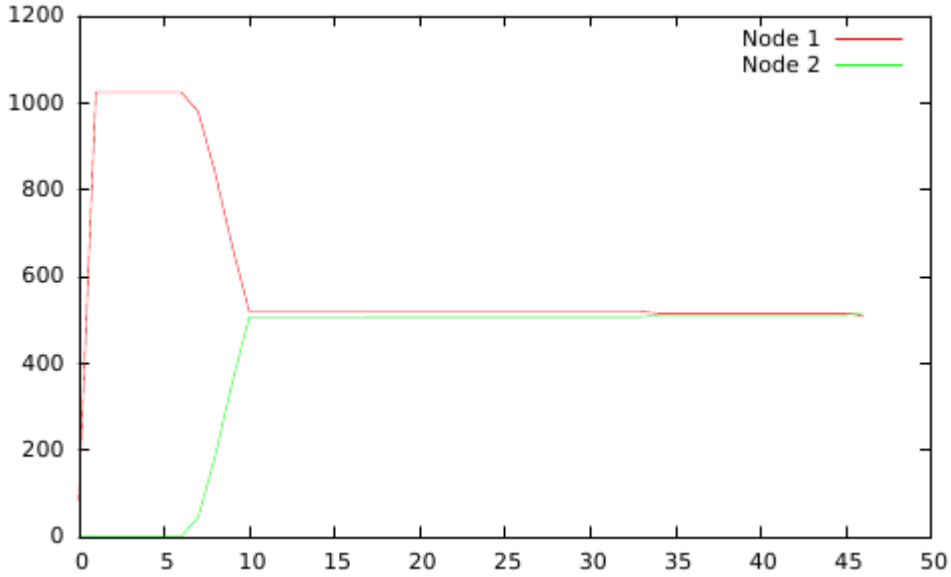
2 nodes

AutoNUMA23 numa02





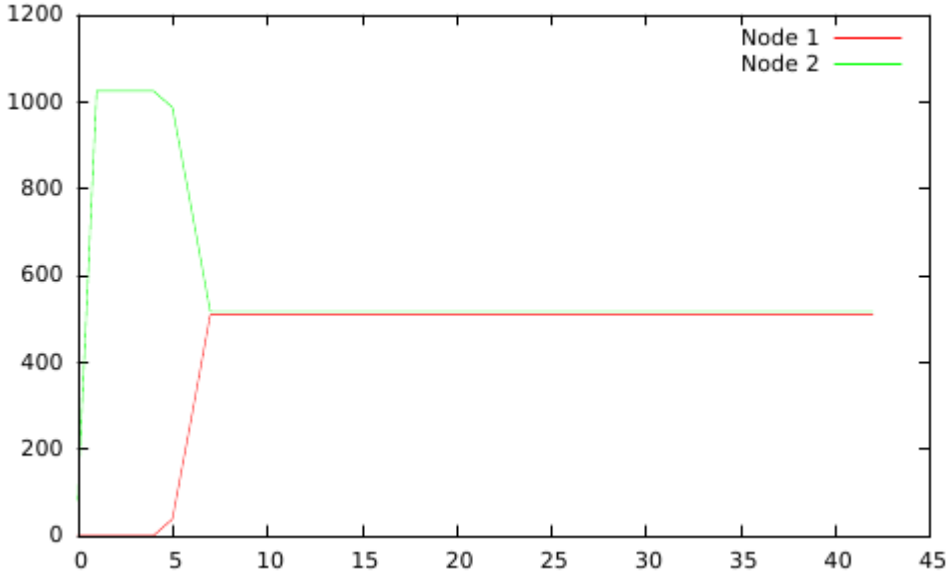


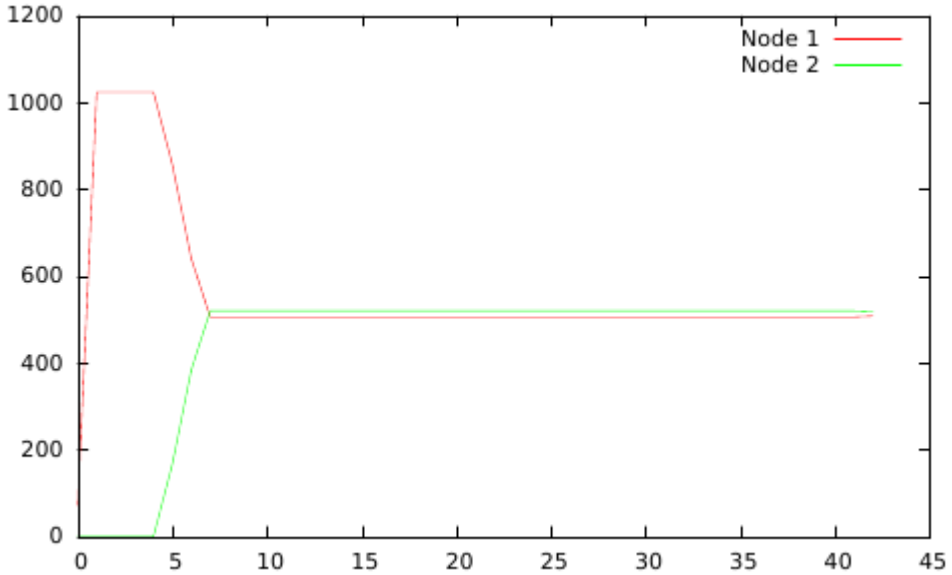


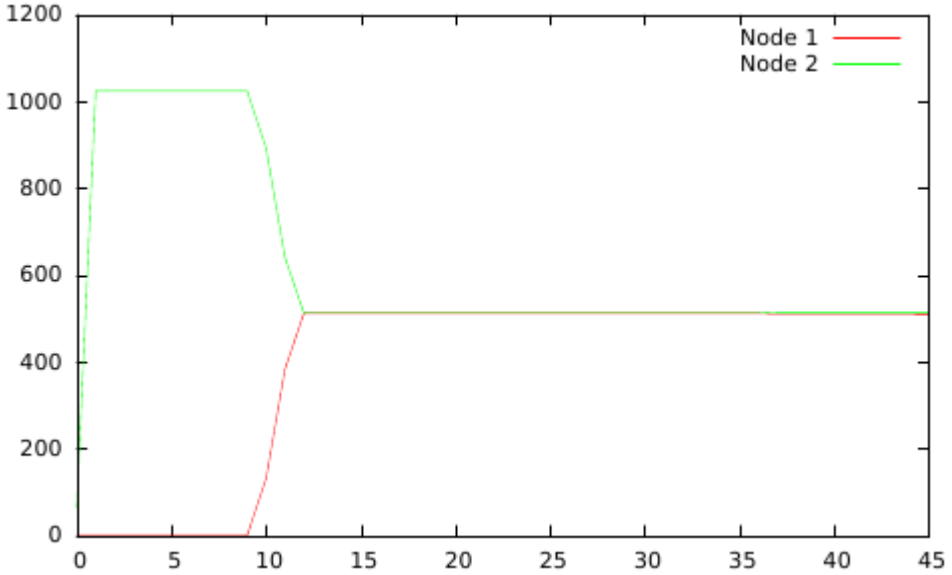
2 nodes

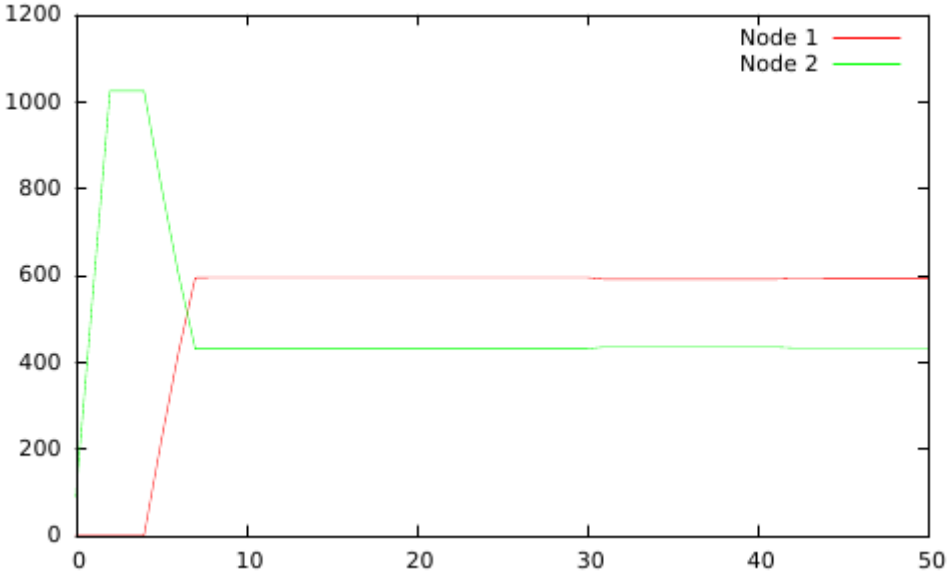
redhat.

AutoNUMA23 numa02_SMT





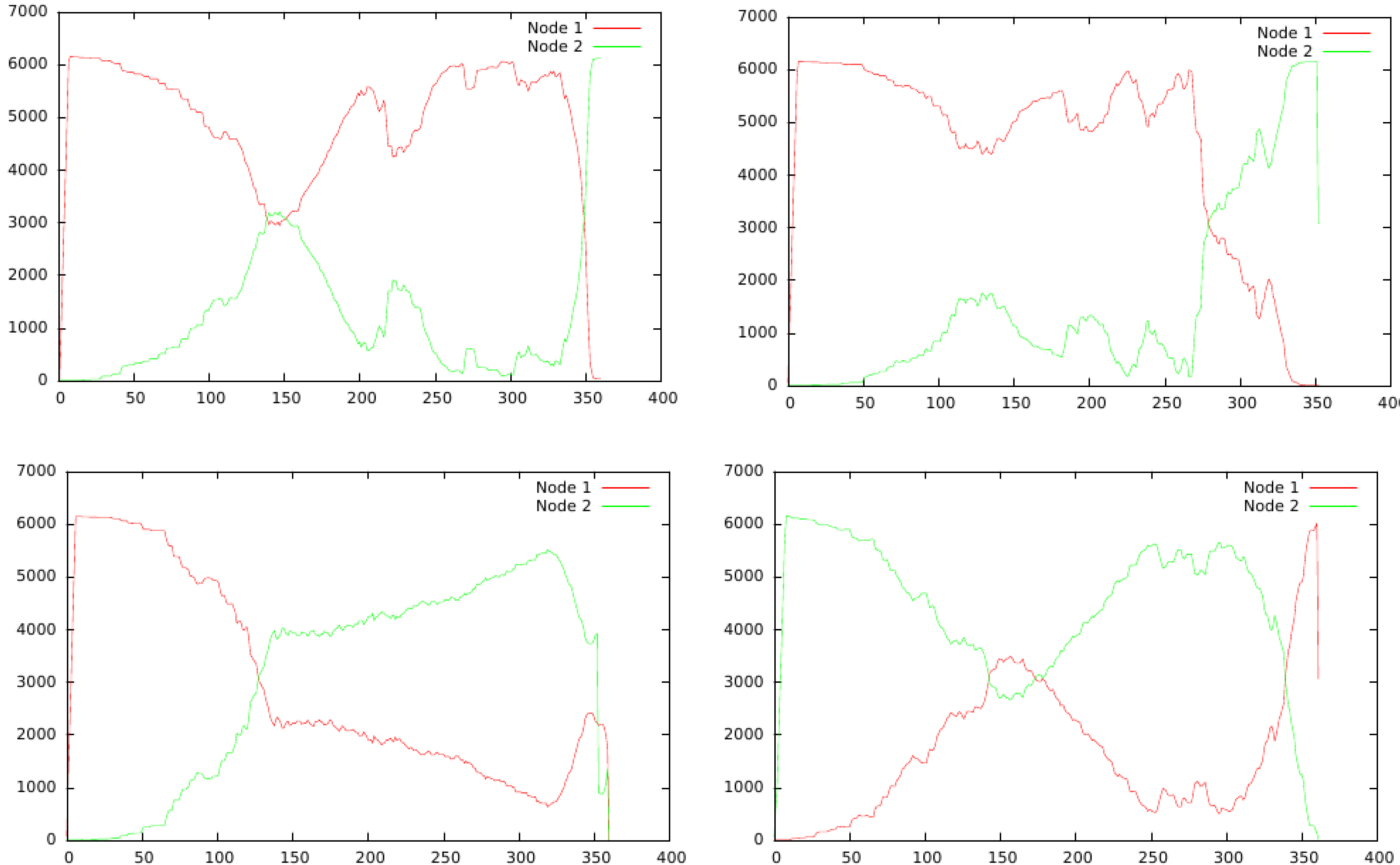




2 nodes

redhat.

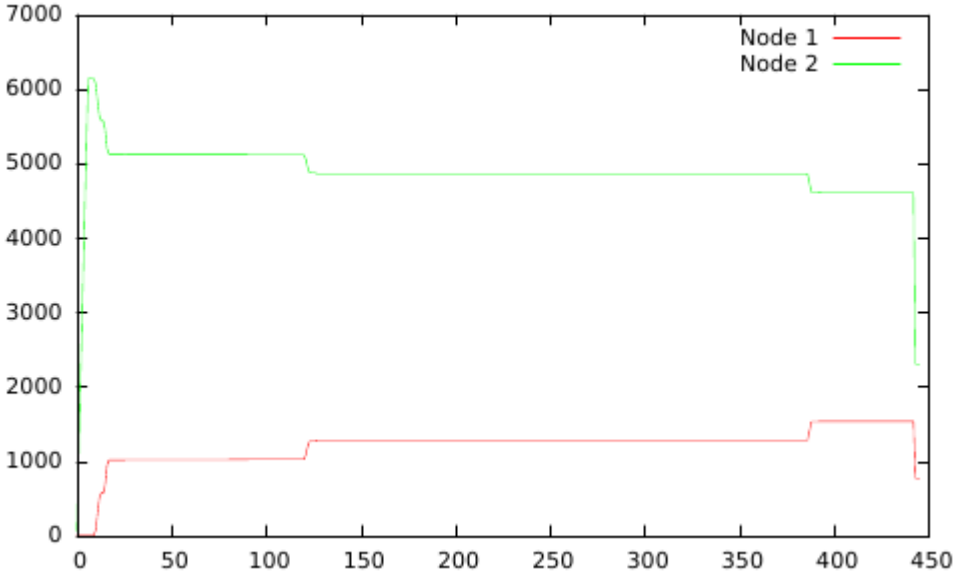
sched-numa-rewrite numa01

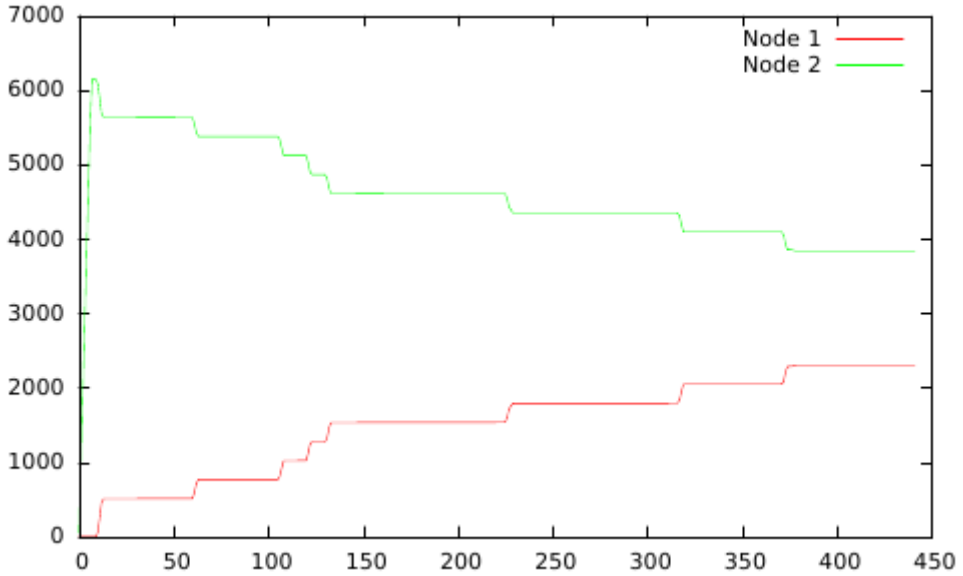


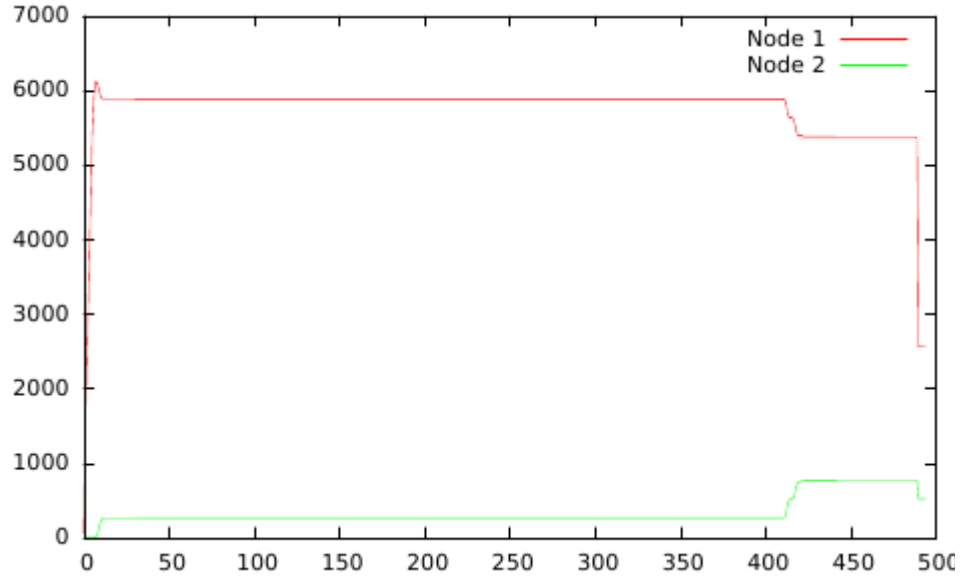
2 nodes

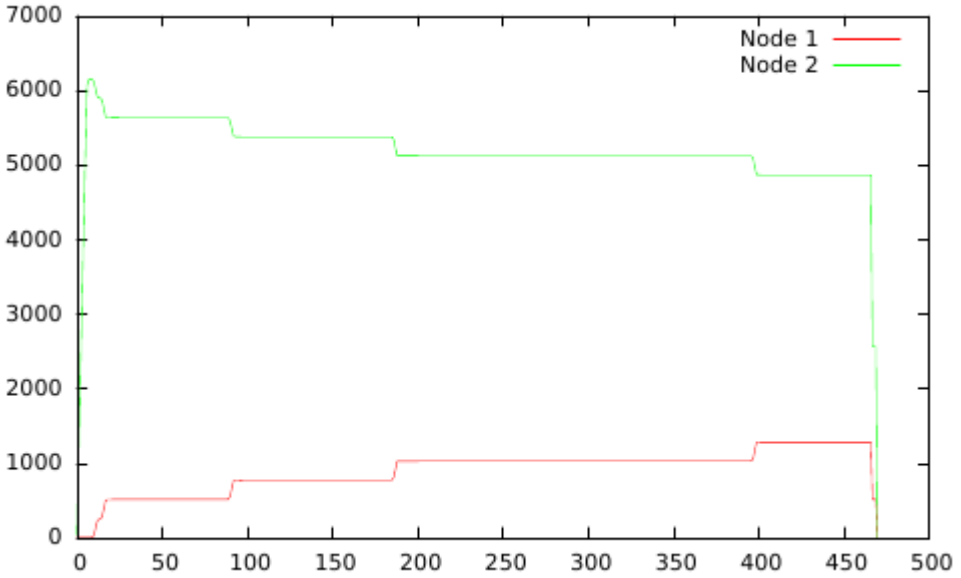
redhat

sched-numa-rewrite numa01_TH..





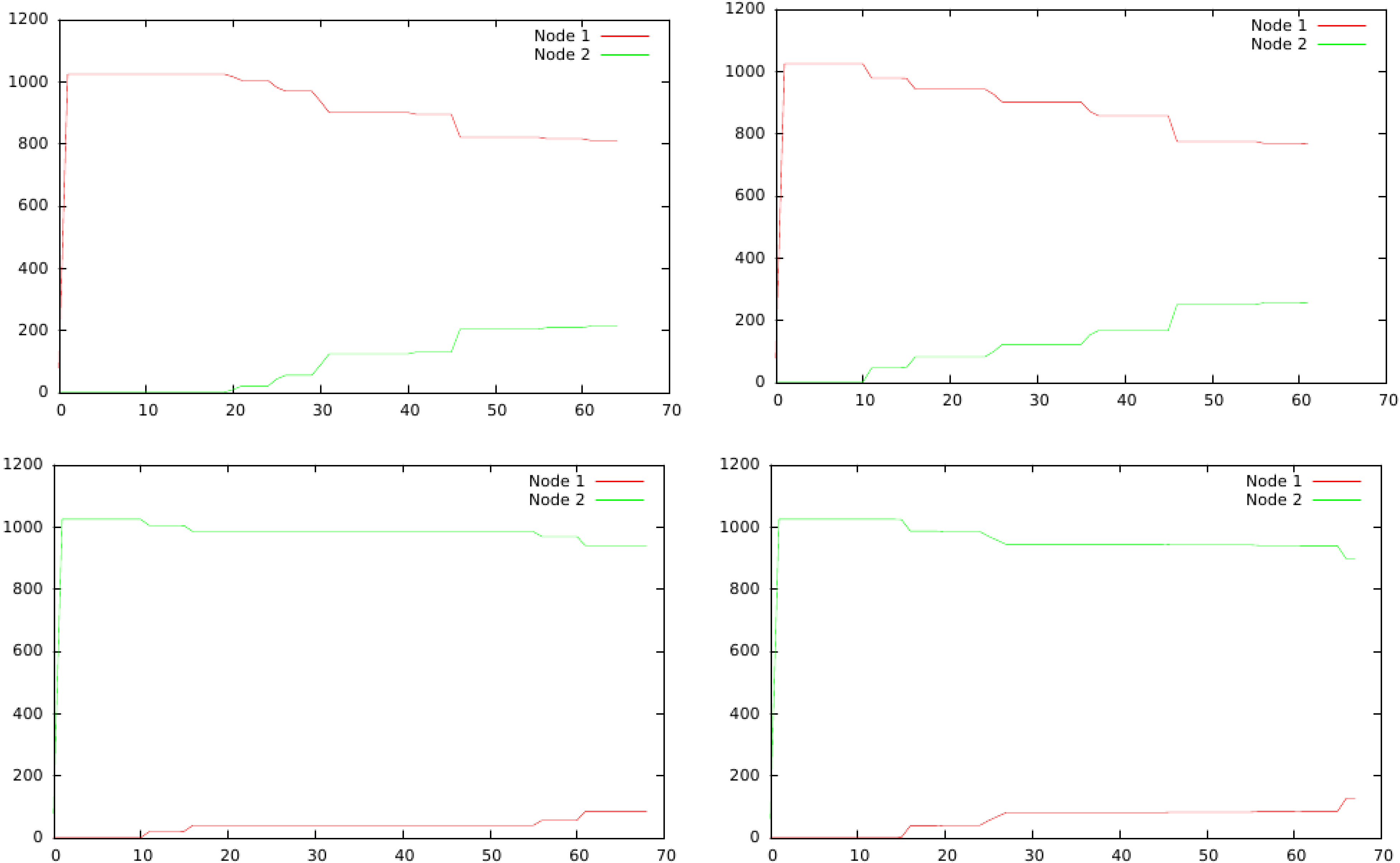




2 nodes

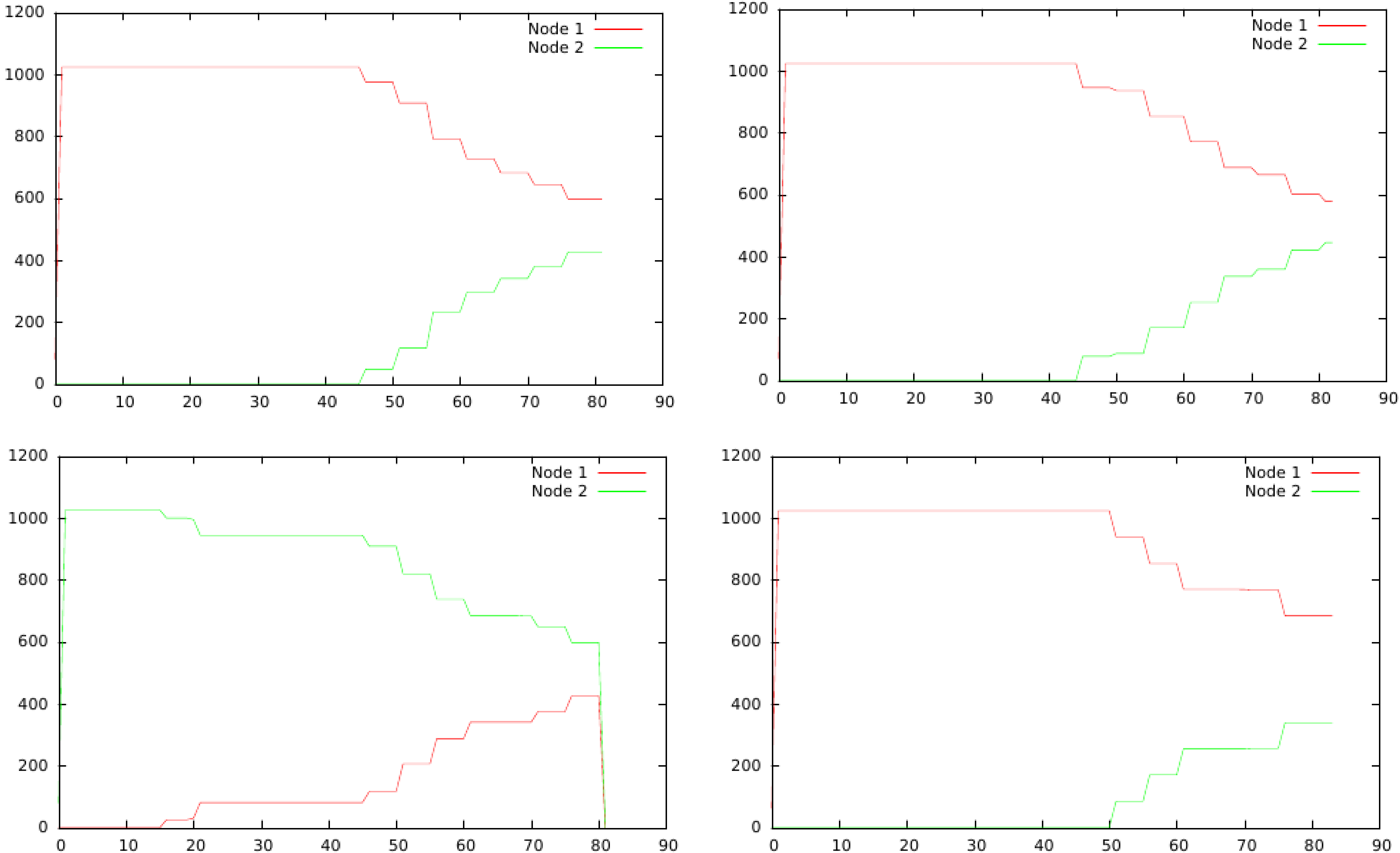
redhat.

sched-numa-rewrite numa02



2 nodes

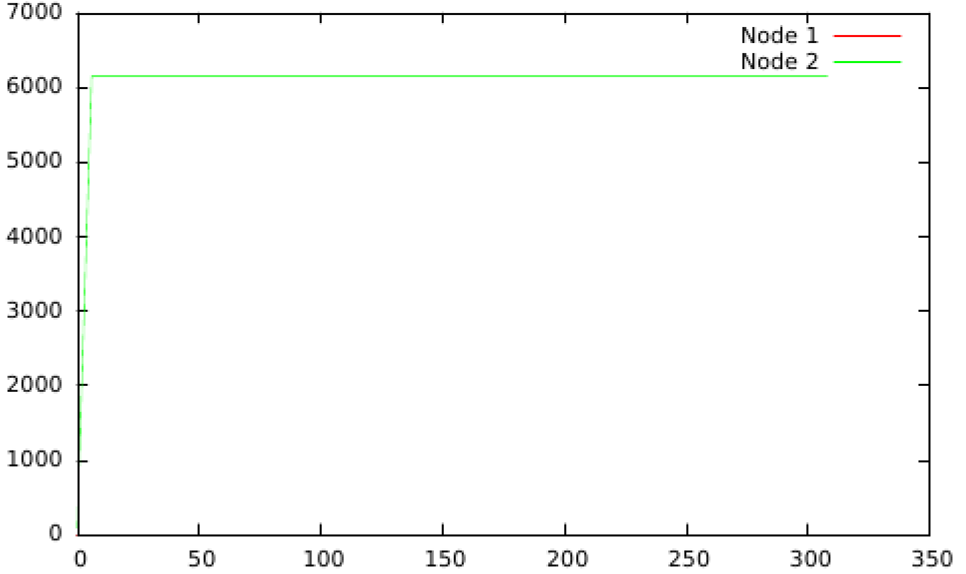
sched-numa-rewrite numa02_SMT

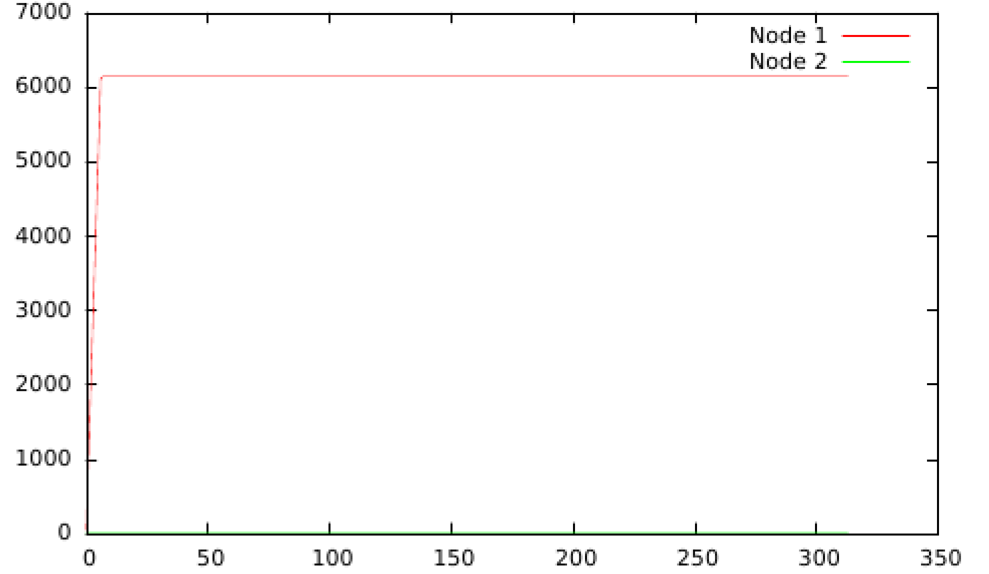


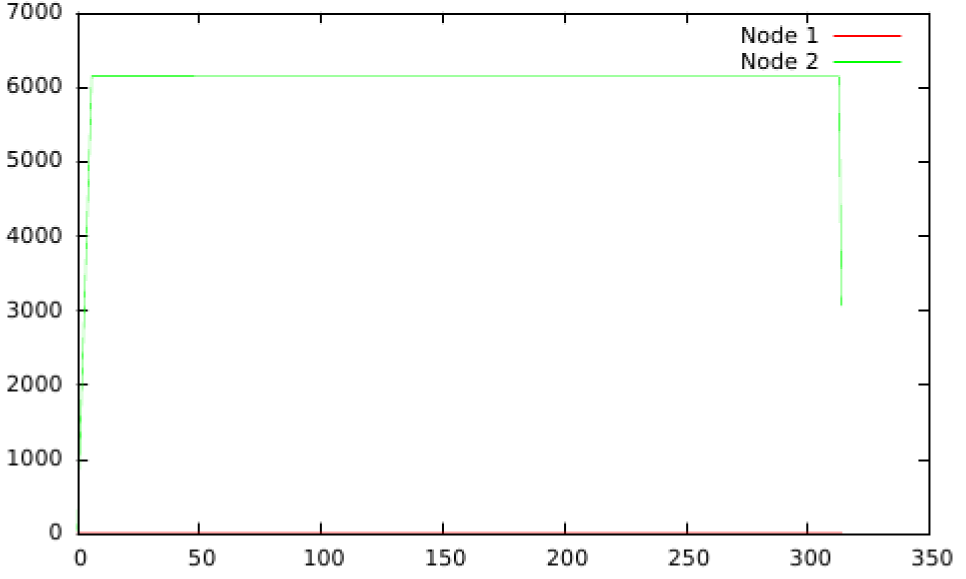
2 nodes

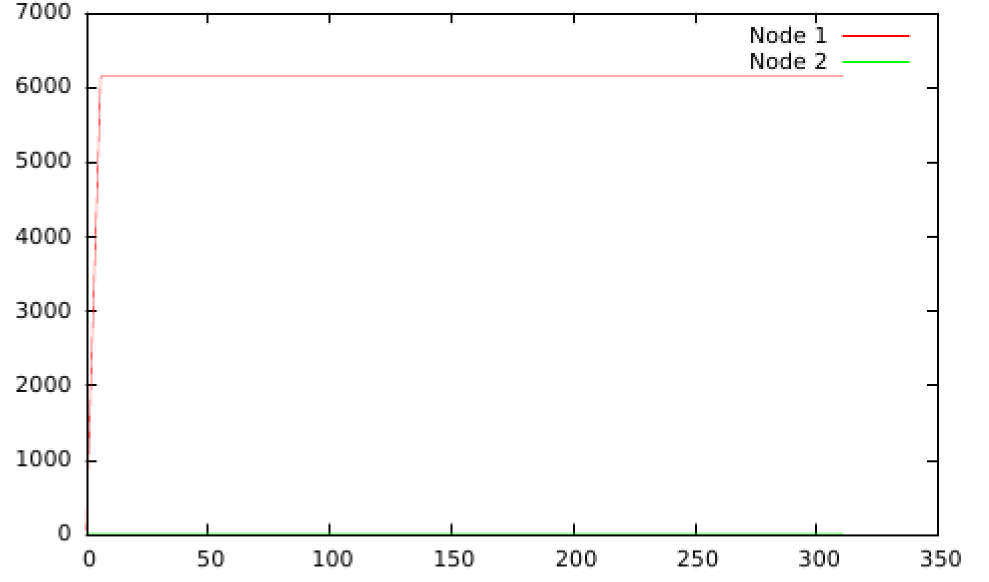
redhat

3.6-rc1 numa01



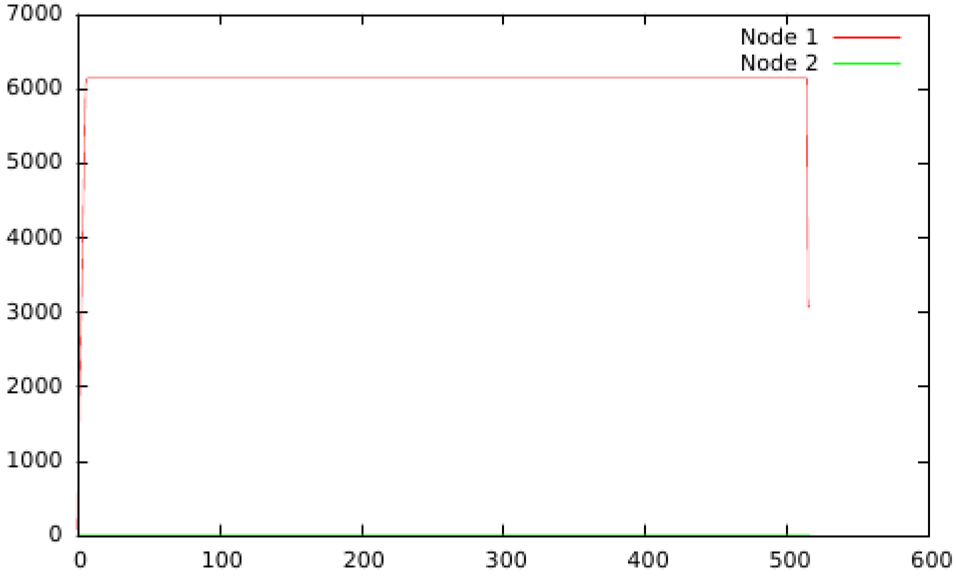


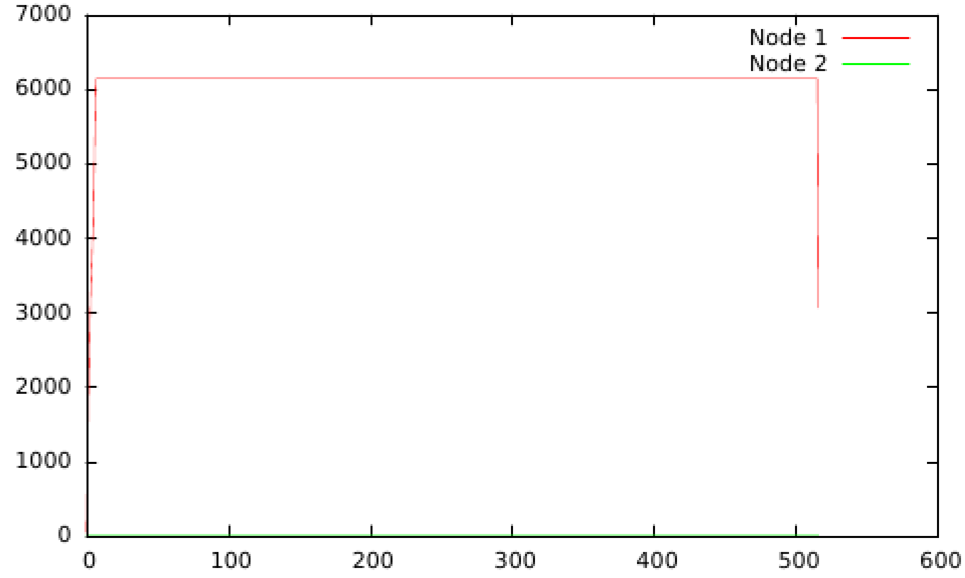


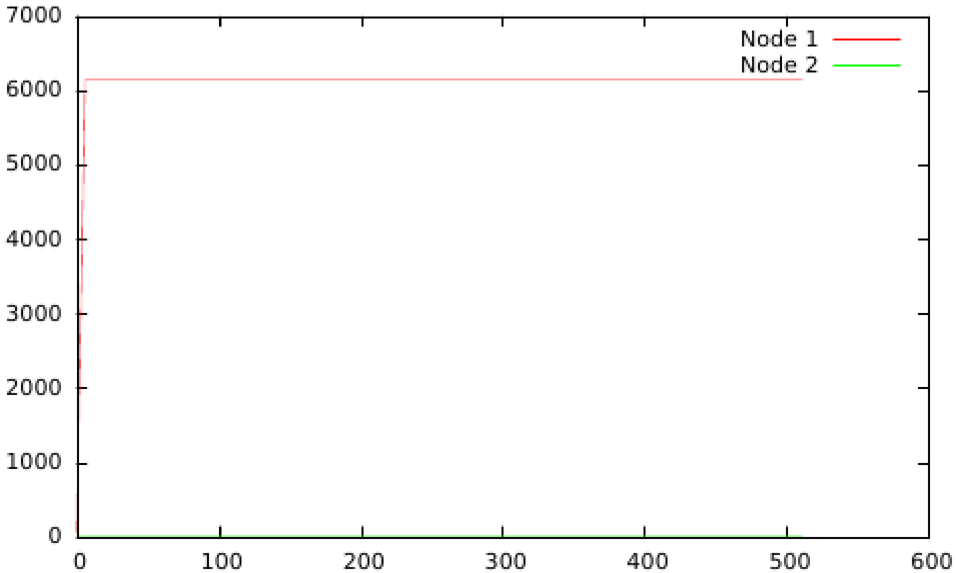


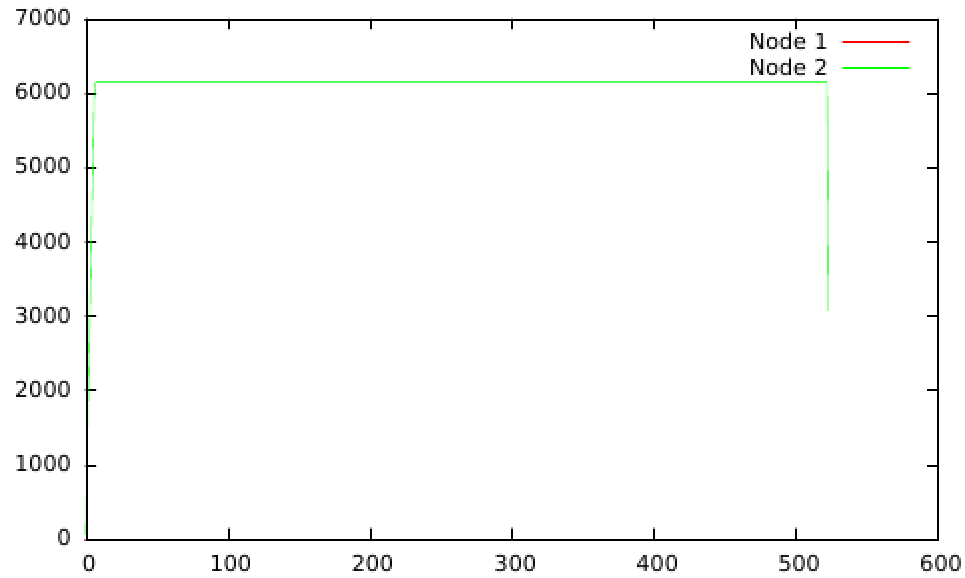
2 nodes

3.6-rc1 numa01_THREAD_ALLOC





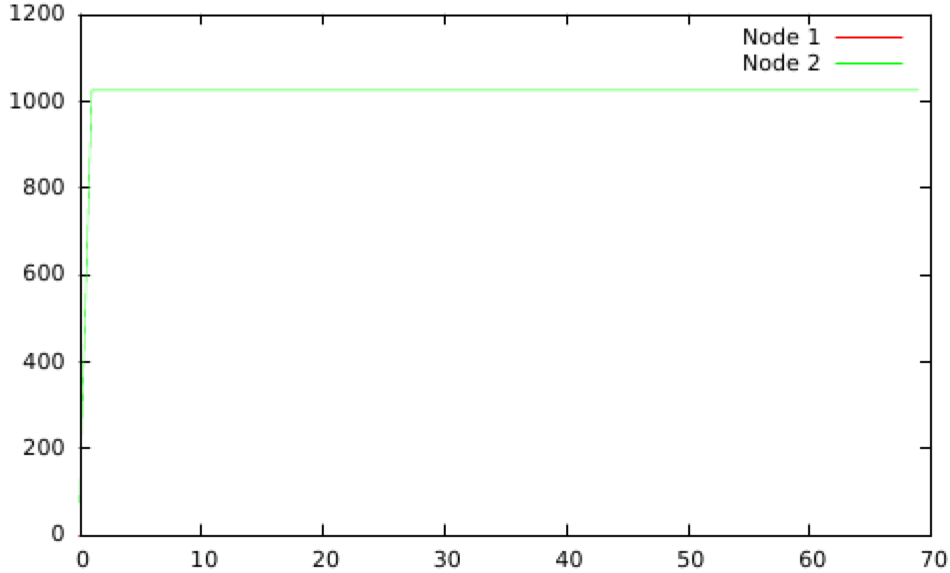


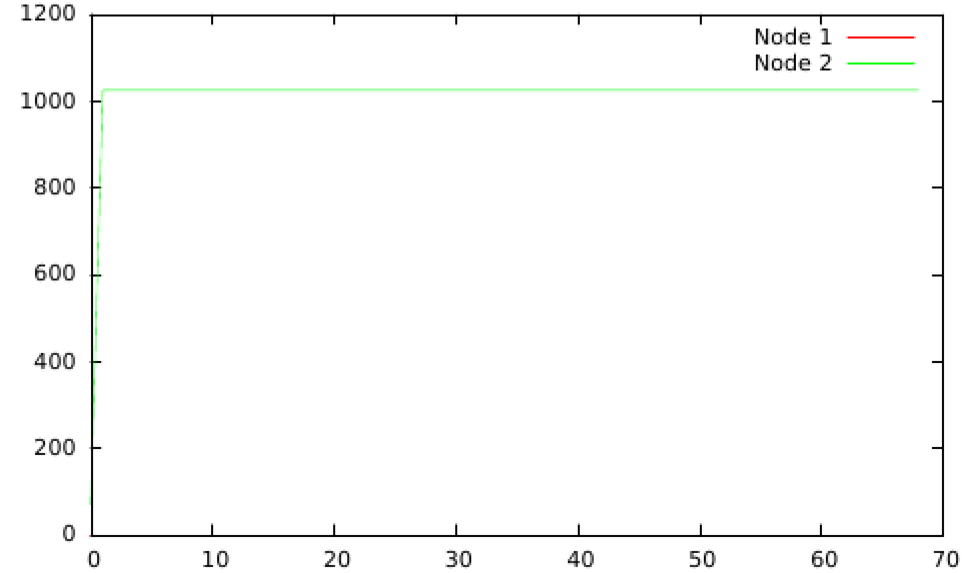


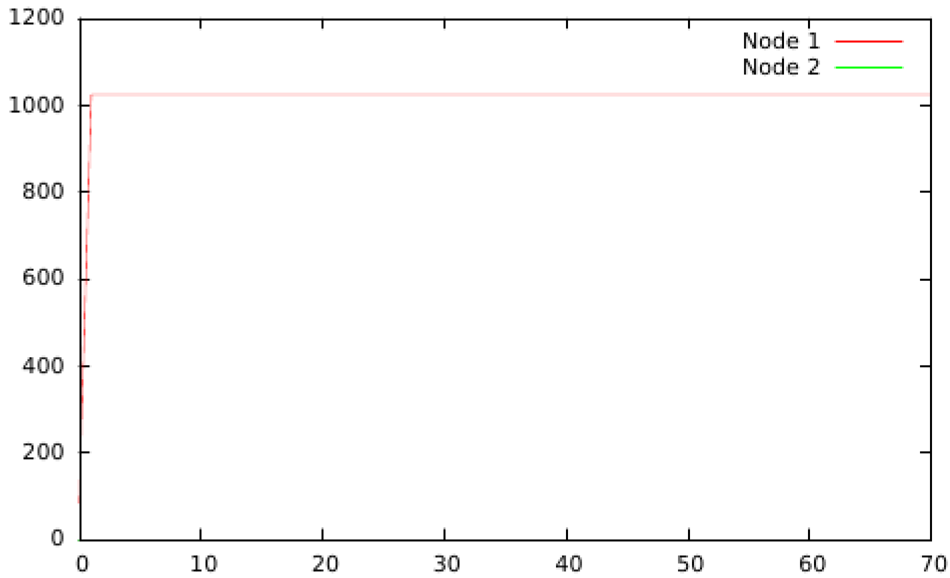
2 nodes

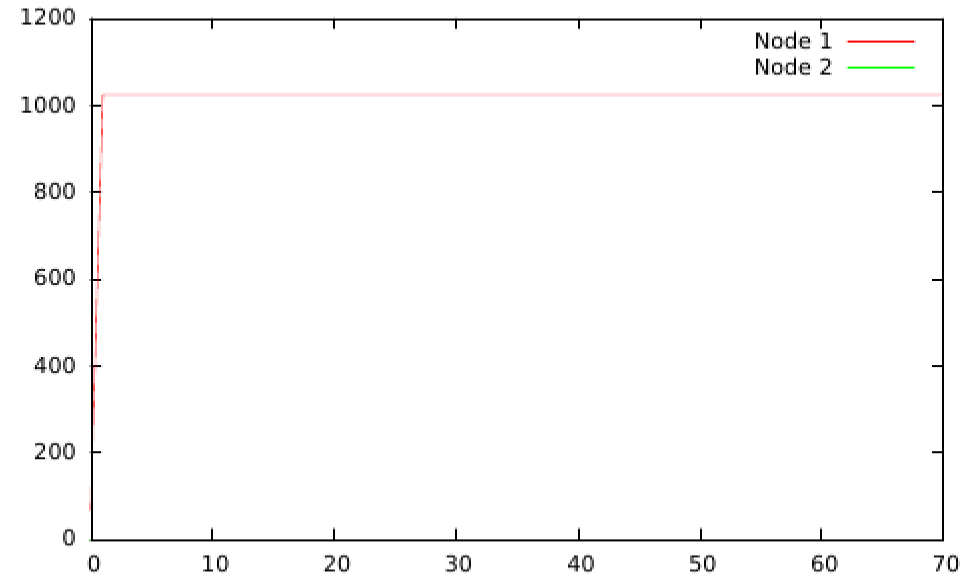
redhat

3.6-rc1 numa02



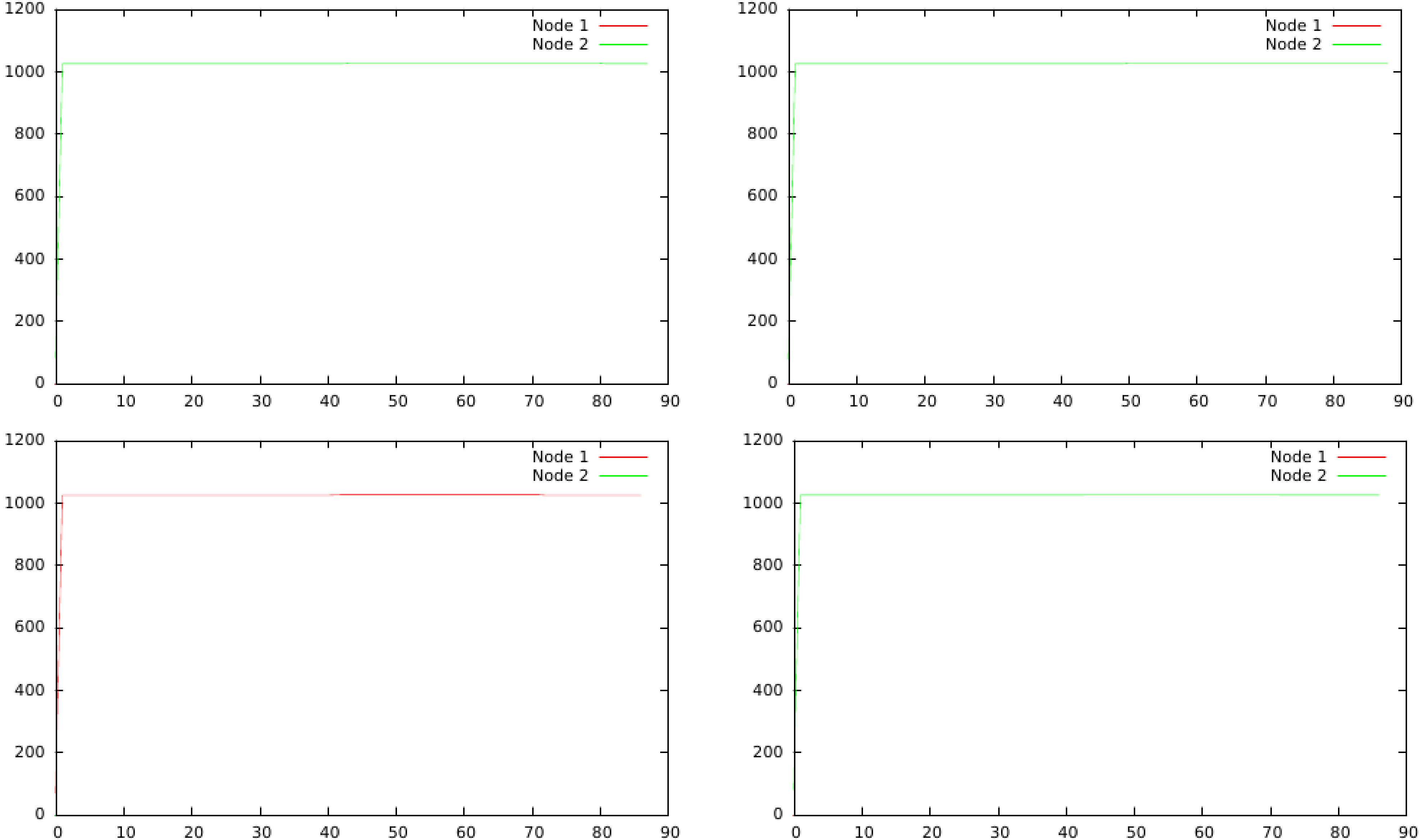






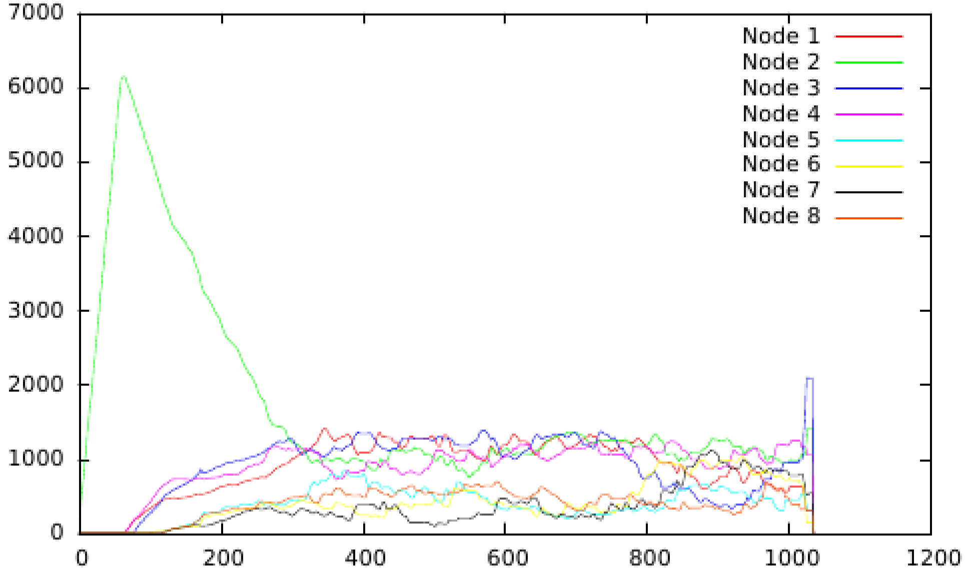
2 nodes

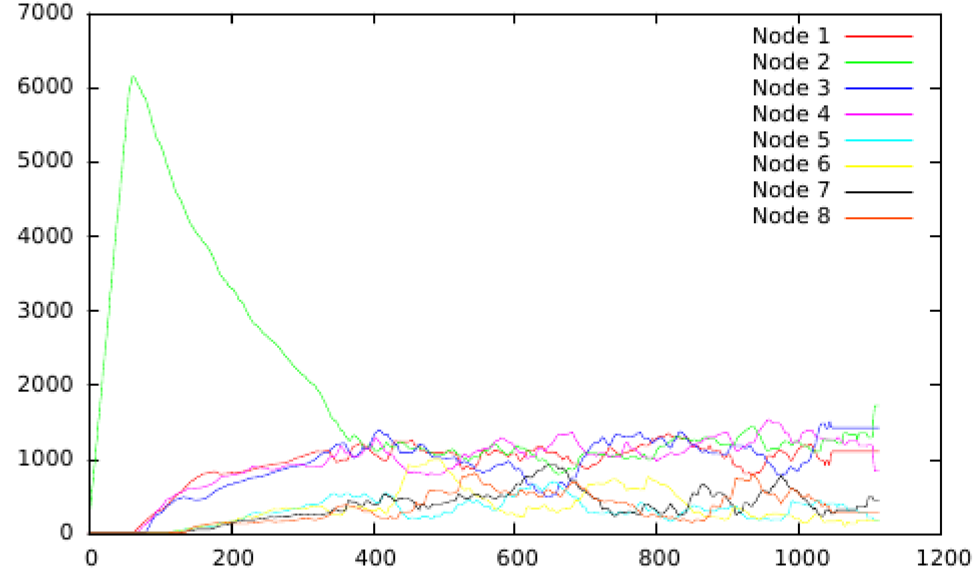
3.6-rc1 numa02_SMT

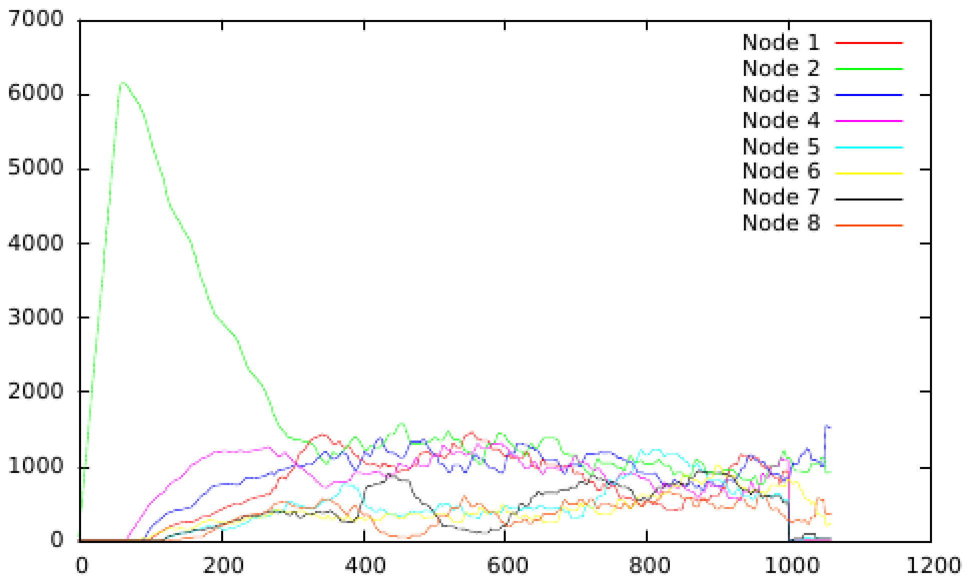


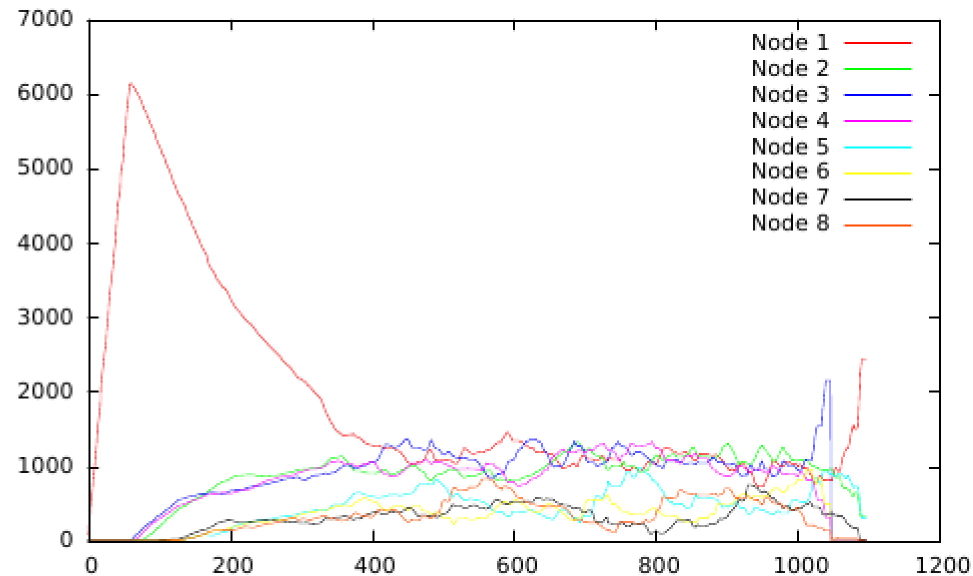
2 nodes

AutoNUMA23 numa01



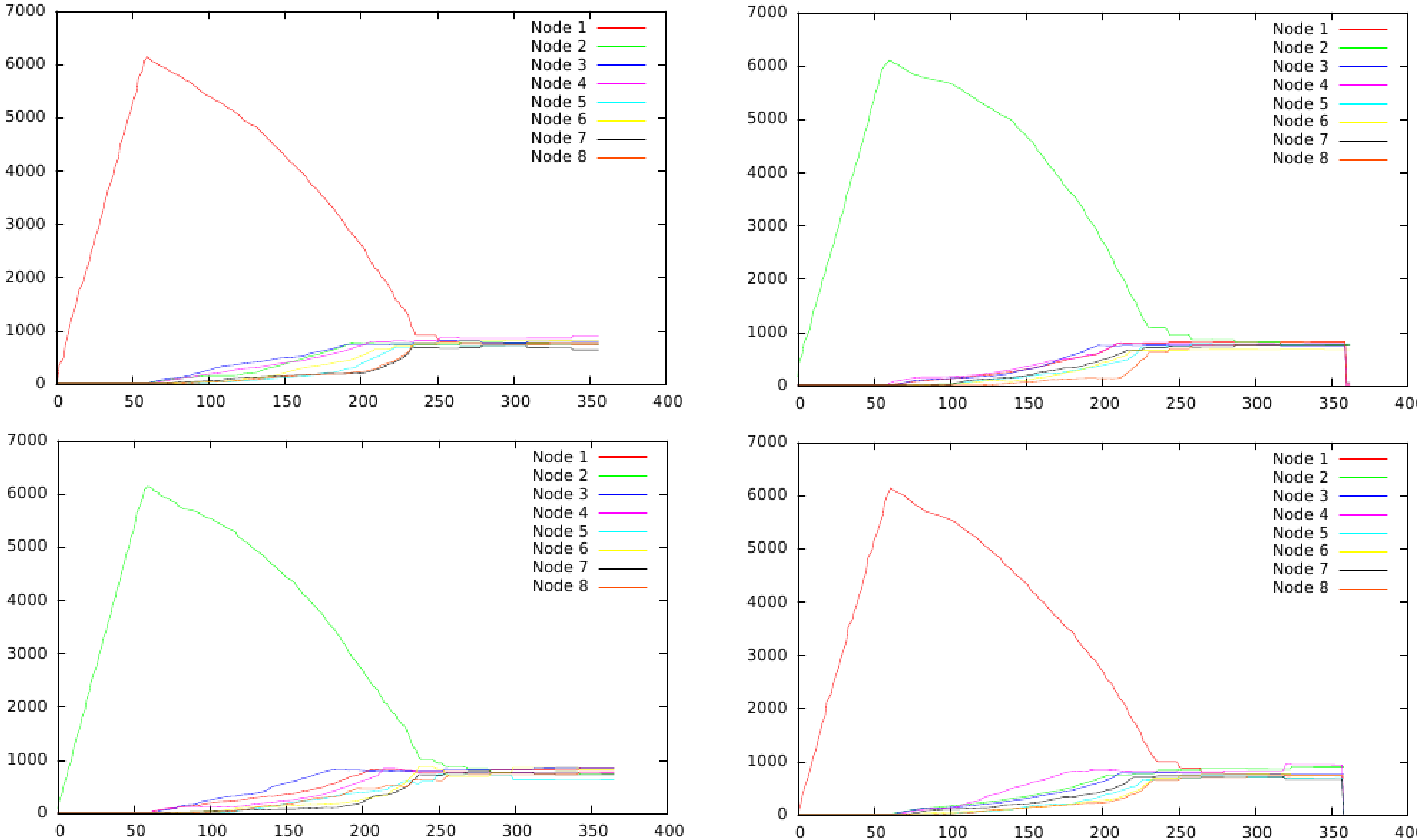






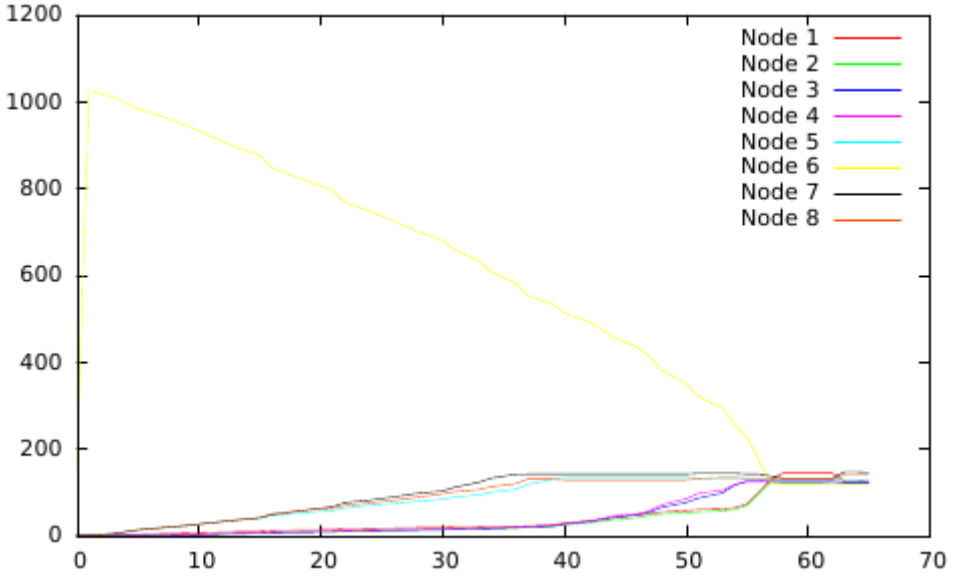
8 nodes → heavy false sharing

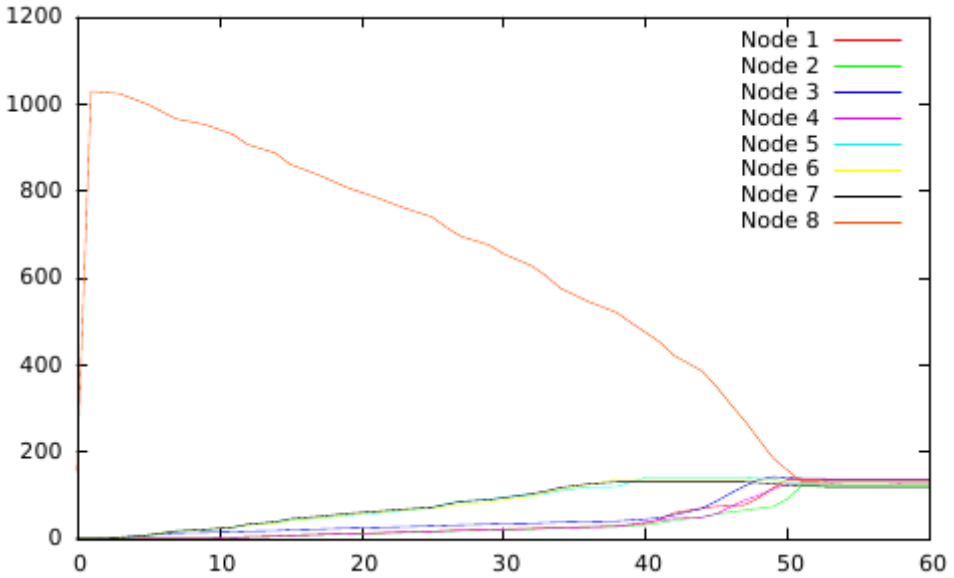
AutoNUMA23 numa01_THREAD..

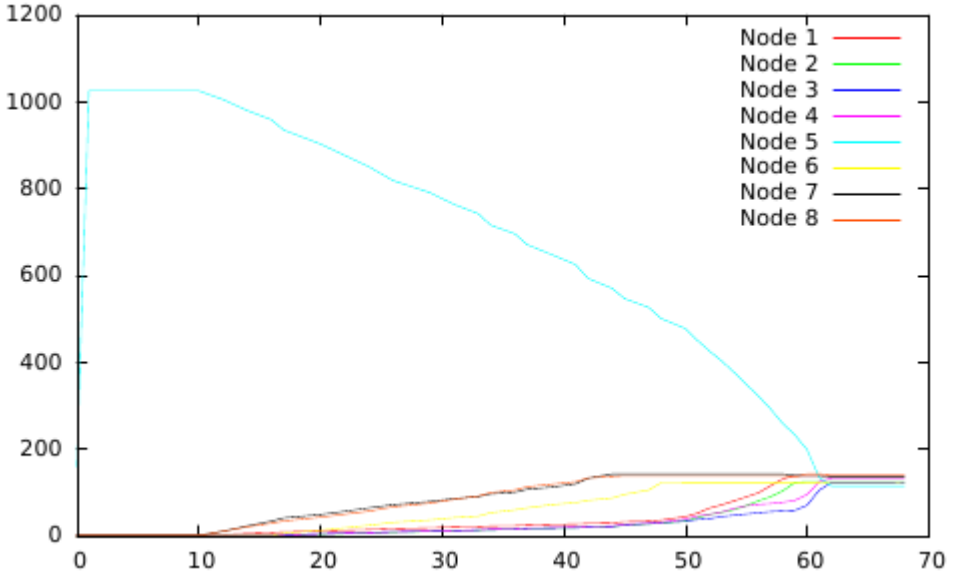


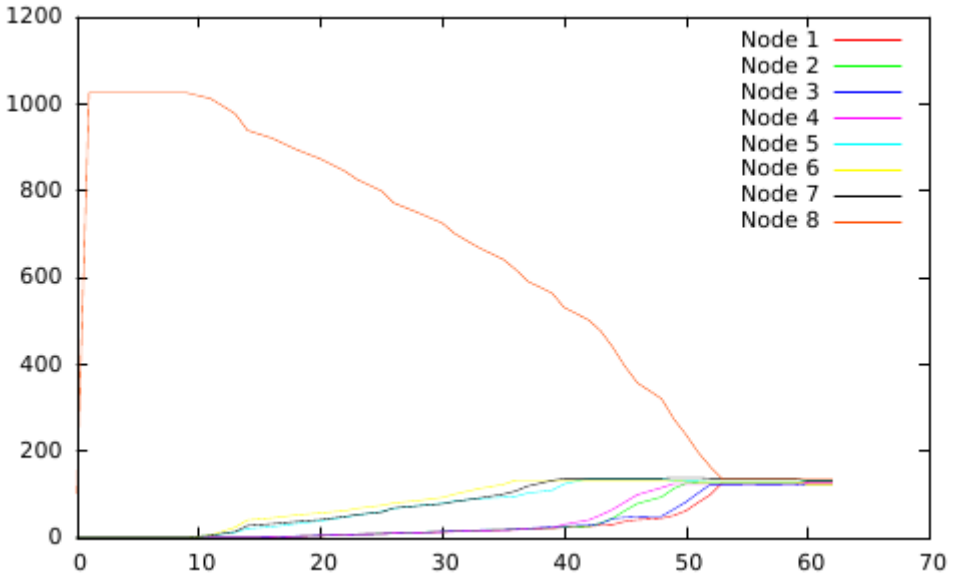
8 nodes

AutoNUMA23 numa02





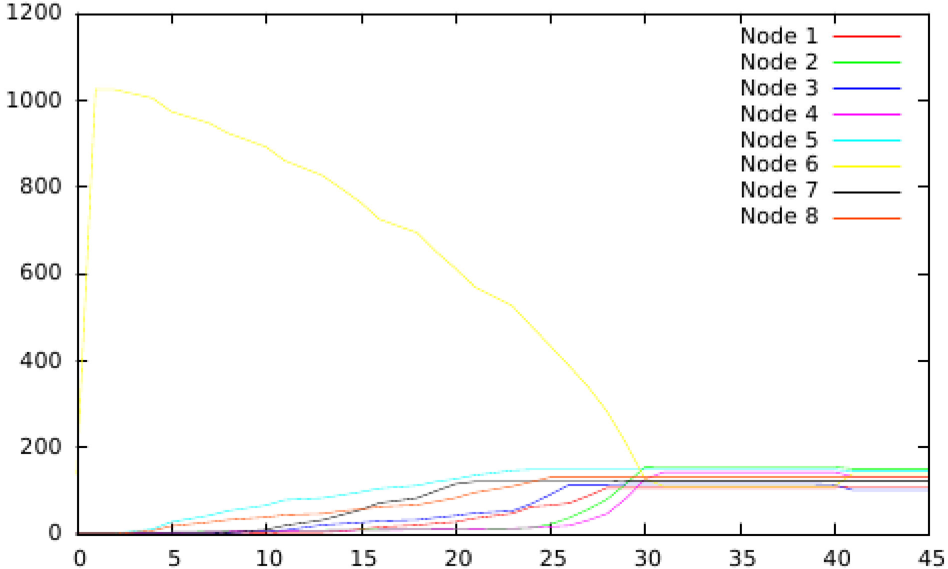


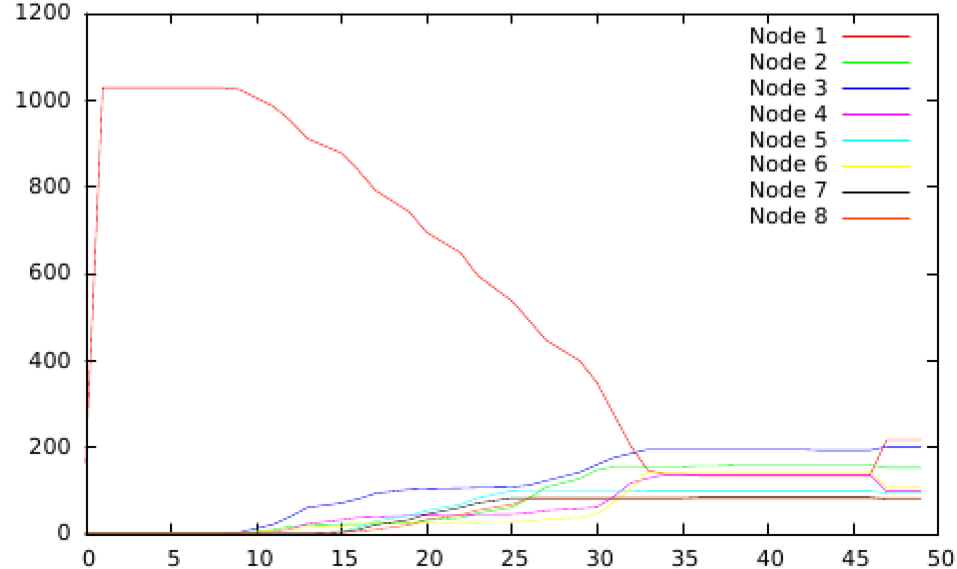


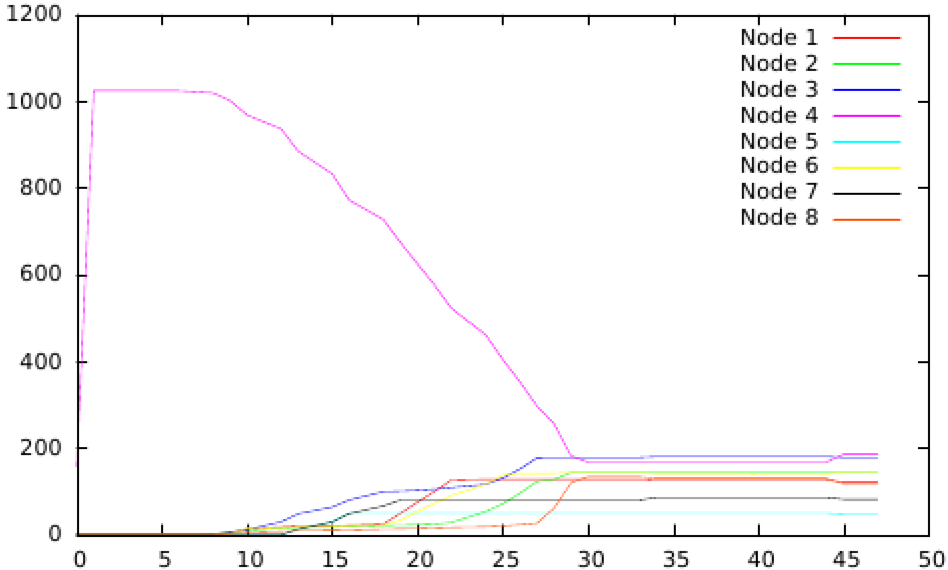
8 nodes

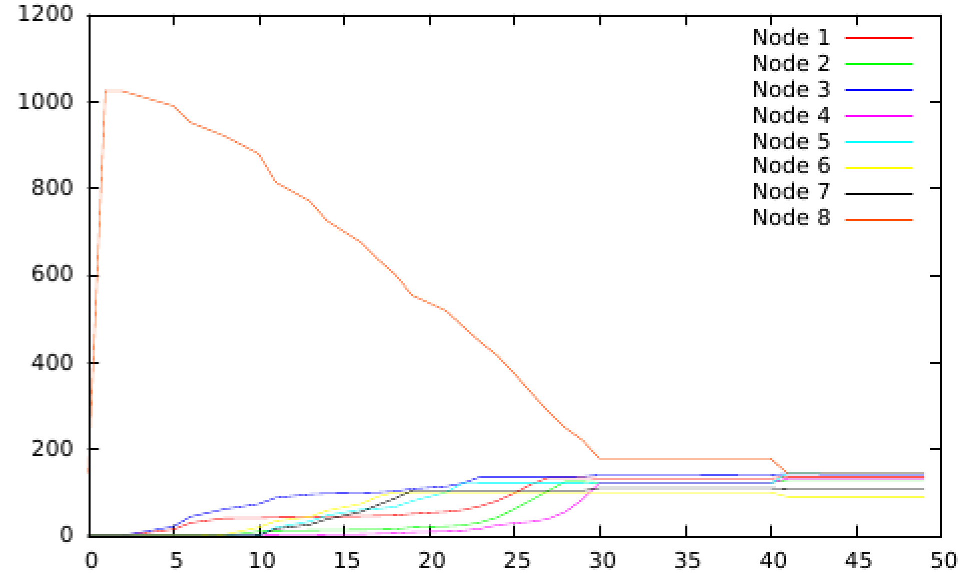
redhat.

AutoNUMA23 numa02_SMT



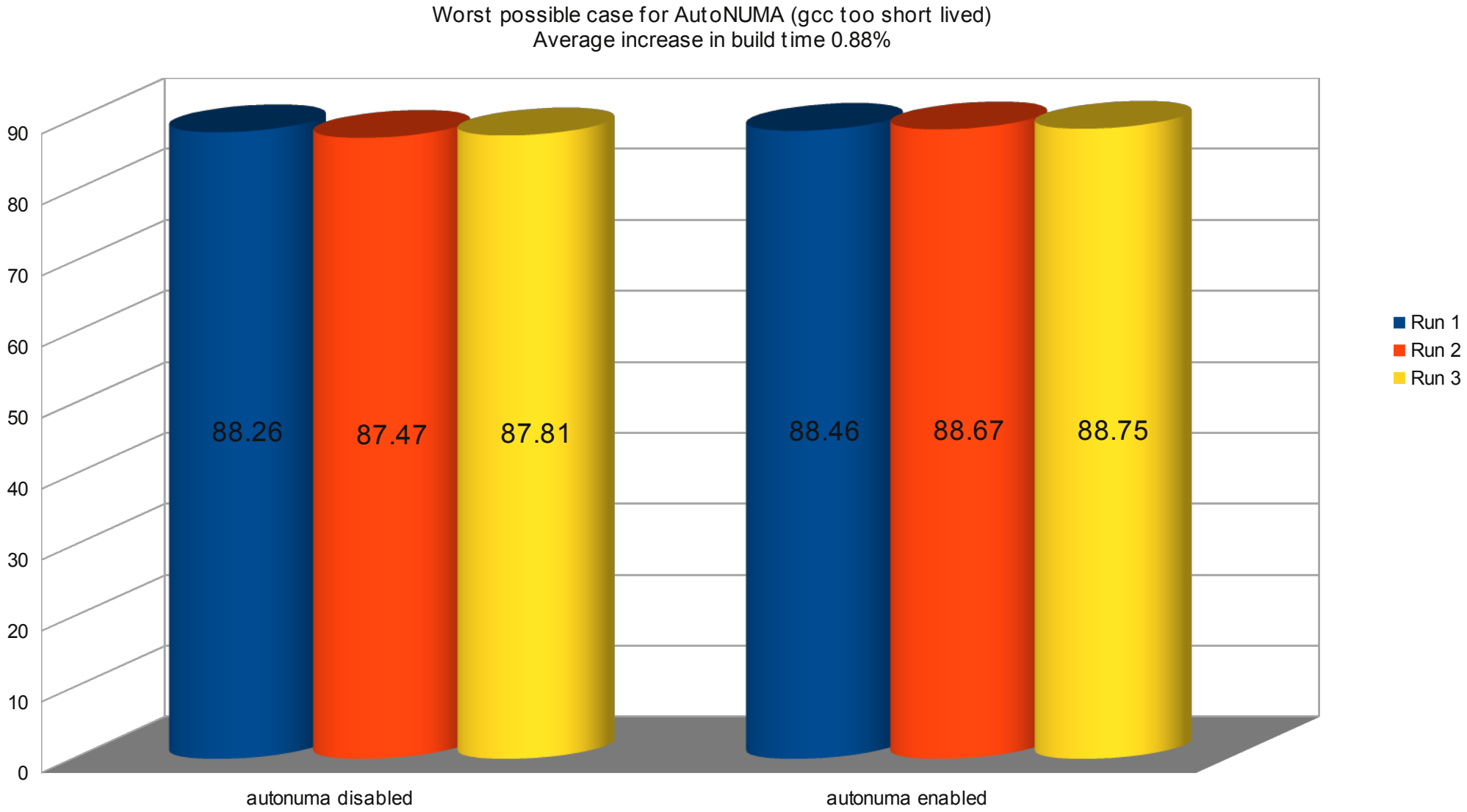






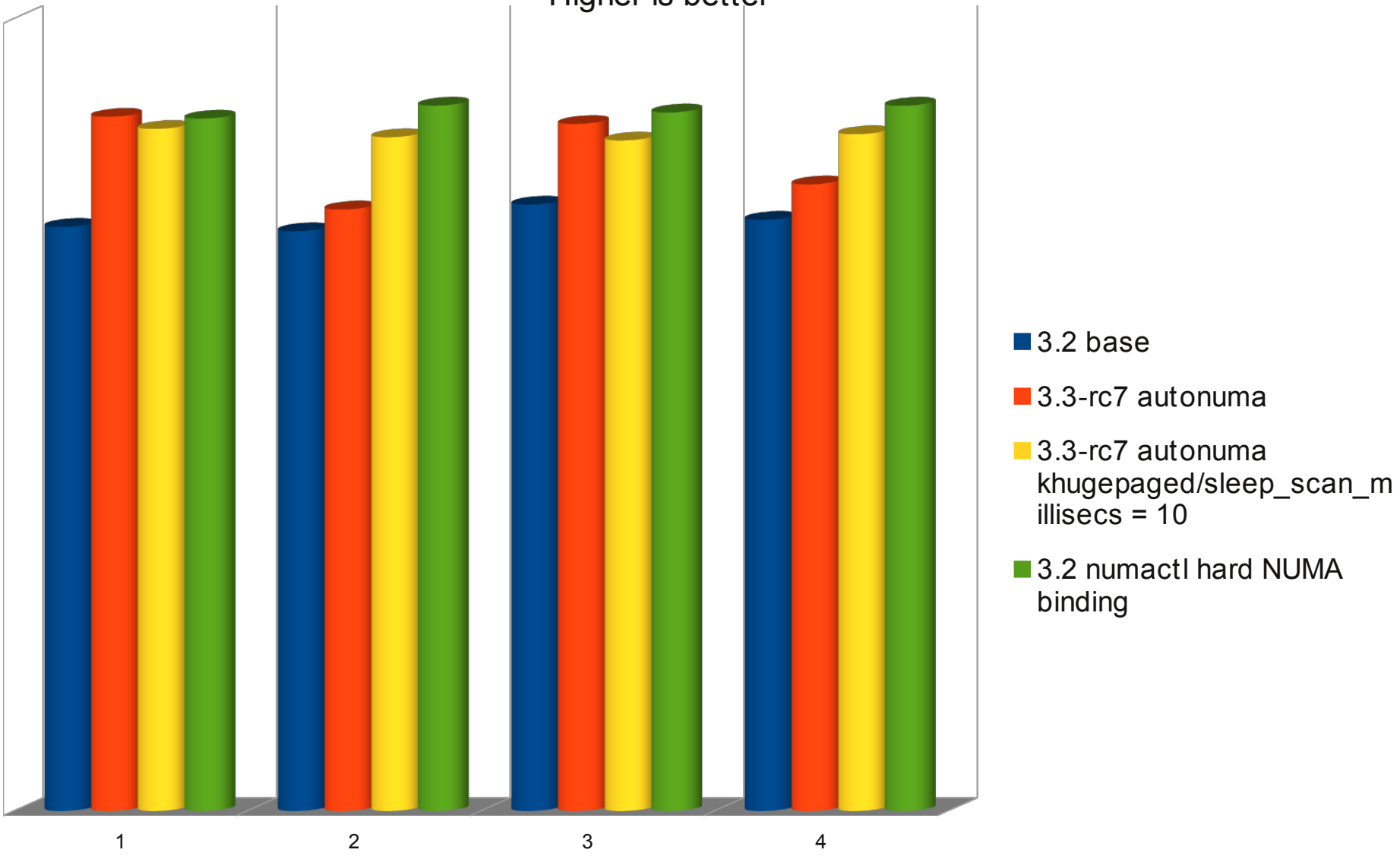
8 nodes

Kernel build time in seconds on tmpfs (make -j32)
Autonuma enabled includes one knuma_scand pass every 10sec
Lower is better



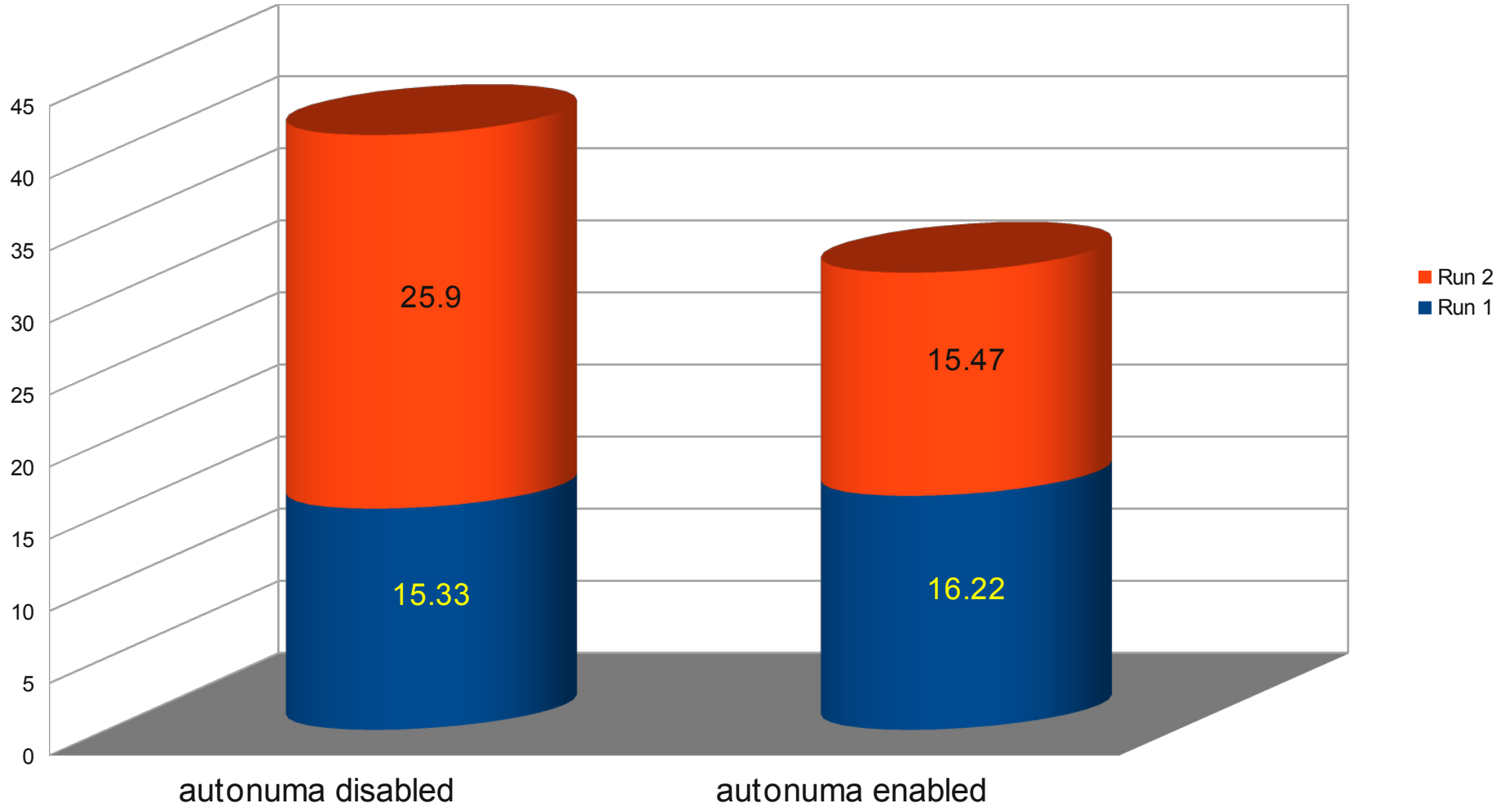
redhat

SPECjbb results 2 NUMA nodes, 8 CPUs per node, 16 CPUs total
THP enabled, no virt
Higher is better



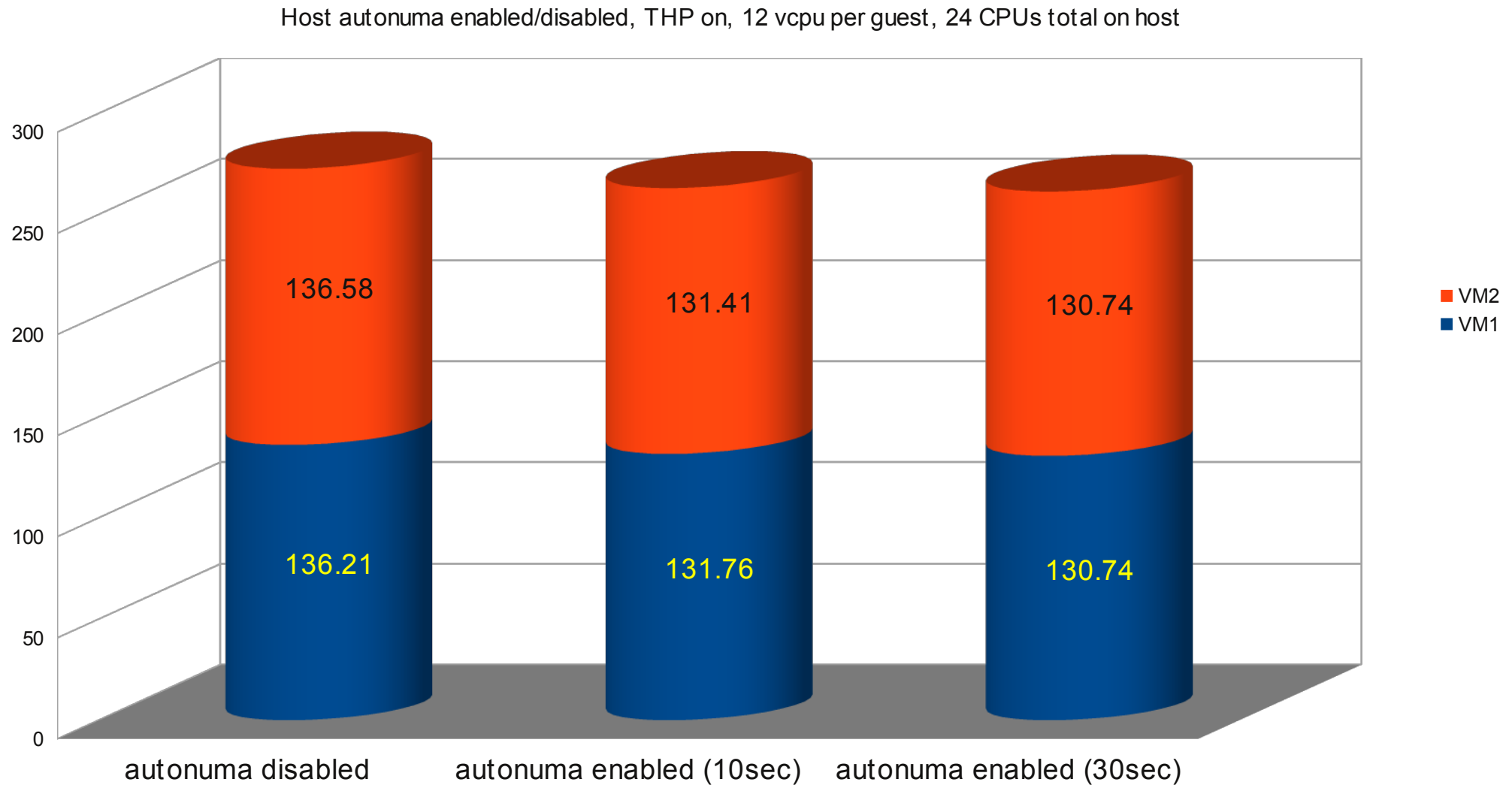
redhat

Virt guest "memhog -r100 1g" (autonuma includes 1 knuma_scand pass every 10 sec)
KVM host autonuma enabled/disabled, THP enabled
Guest VM fits in one host NUMA node
Lower is better



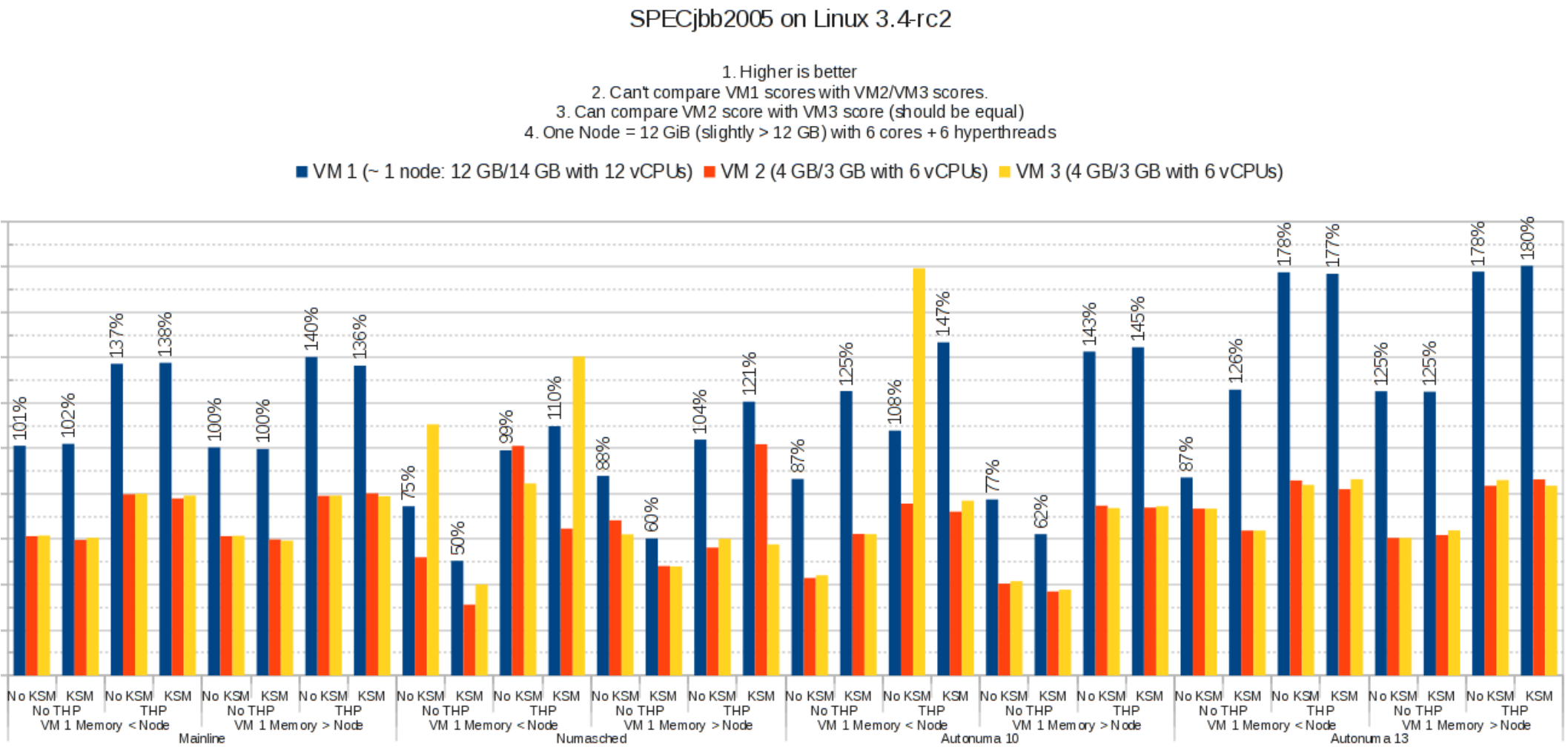
redhat

kernel build -j16 in parallel in 2 KVM (both in tmpfs, in a loop started in sync)
Both guest VM fits in one host NUMA node
autonuma/knuma_scand/scan_sleep_pass_millisecs = 5000 | 15000 (10sec | 30sec)
Lower is better



redhat

KVM/KSM/THP bench run by IBM



redhat.