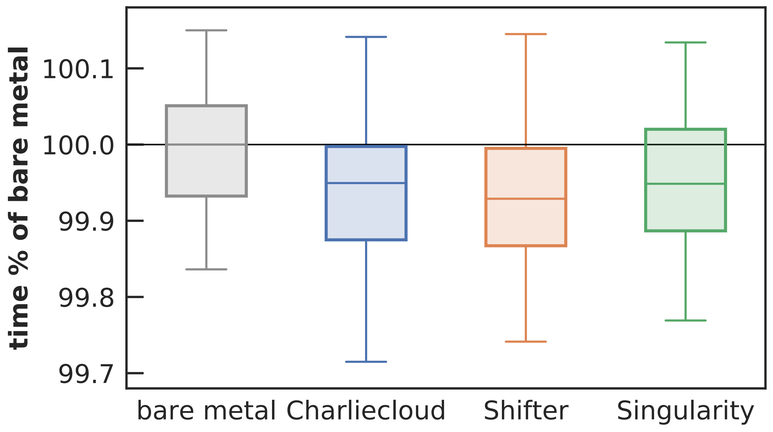
HPC container runtime performance overhead: At first order, there is none

Alfred Torrez, Reid Priedhorsky, Tim Randles
HPC Division, Los Alamos National Laboratory

LA-UR 19-27967

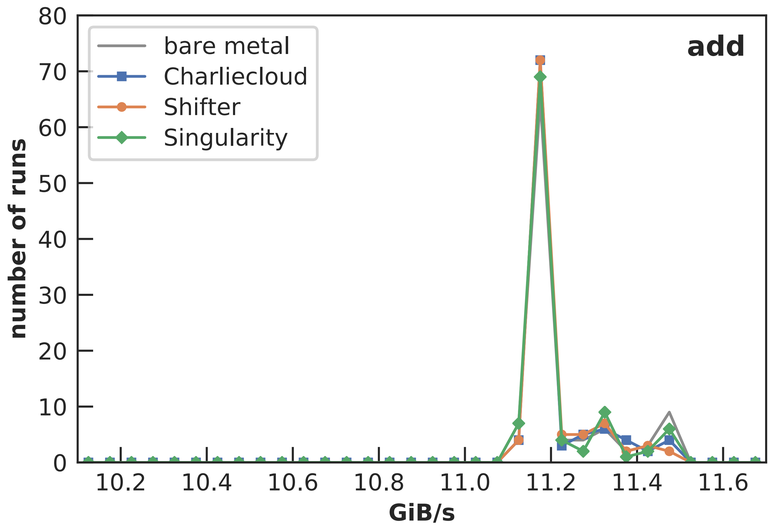
Research question: Do containers introduce performance overhead for HPC applications? Prior work addressed this question one technology at a time. We used industry standard benchmarks SysBench, STREAM, and HPCG to evaluate all three HPC container runtimes: Charliecloud, Shifter, and Singularity. We found no significant performance impact and a modest memory overhead.

CPU performance



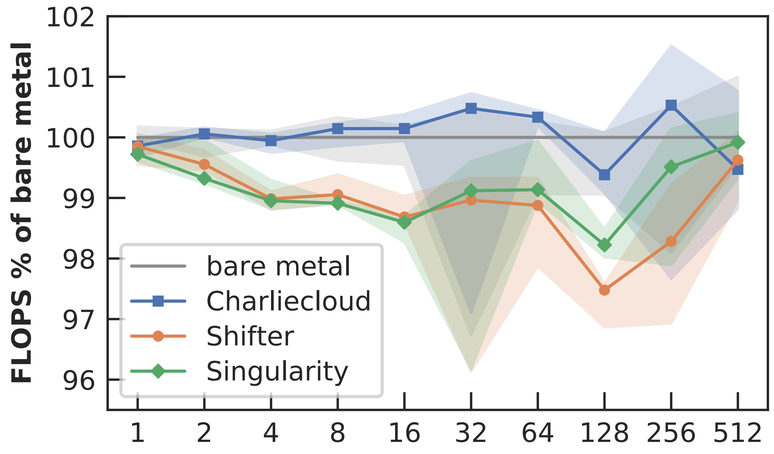
SysBench prime-number computation time relative to bare metal's 129.36s. The four technologies have essentially the same performance; the spread across all 100 runs/technology was only 0.4%.

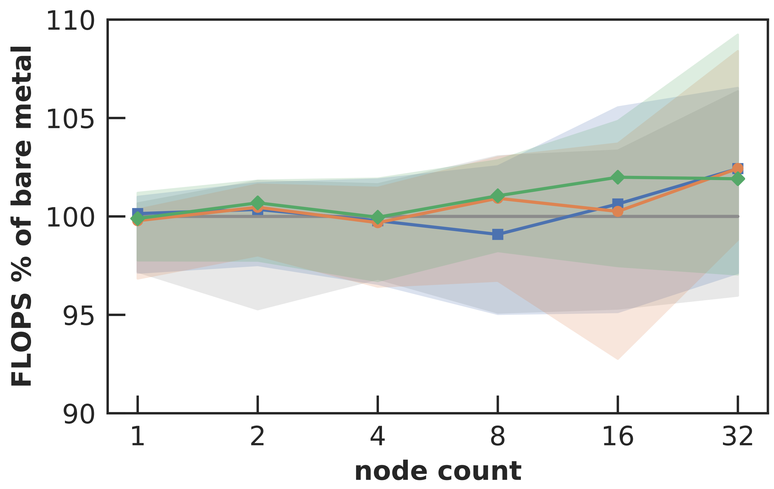
Memory performance



STREAM memory bandwidth. Performance on "add" and the other three kernels tested (100 runs/kernel/technology) is again essentially identical.

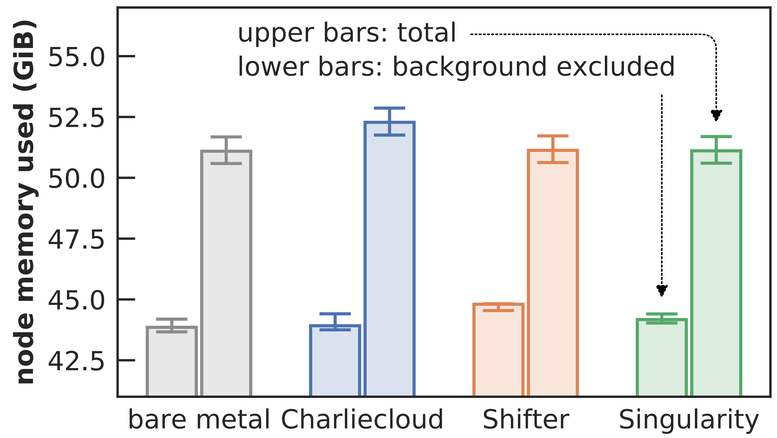
Application performance



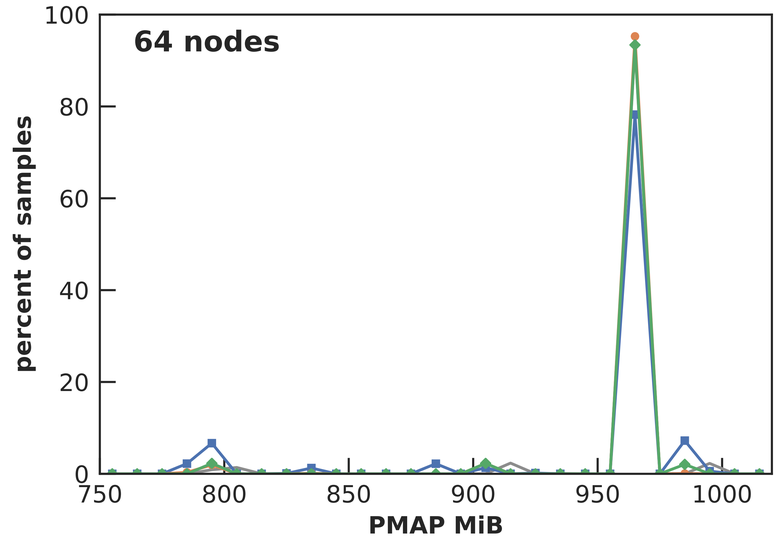


HPCG performance relative to bare metal. The upper plot shows results in our scaling test with 4–10 runs per condition. Performance of the container technologies was close to bare metal, but not as close as the other tests. The lower plot shows 50 runs per condition. The container technologies cluster together closely, suggesting that the two clusters in the upper plot are an artifact. We suspect that the apparent differential between bare metal and containers at 32 nodes is the results of minor shared library linking differences. These results are again consistent with zero performance difference between any of the four technologies.

Memory usage



STREAM memory usage sampled from /proc/meminfo. Memory use is broadly similar across the four technologies. We suspect that Charliecloud's extra total 1.2 GiB is due to the images being stored in a tmpfs (optional); we do not yet understand Shifter and Singularity's memory overhead with background excluded.



HPCG application memory usage sampled from pmap(1). The four technologies had essentially identical usage here and 1, 8, and 512 nodes.

Hardware. LANL's CTS-1 clusters Grizzly (1490 nodes, 128 GiB RAM/node) and Fog (32 nodes, 256 GiB RAM/node). These systems have 36 CPU cores on an Intel S2600KP motherboard with 2× Intel E5-2695v4 (Broadwell) 2.1 GHz 18-core CPUs; hyperthreading is disabled. Node memory is 16GiB DDR4 2400MHz DIMMs. Cluster interconnect is Intel OmniPath OP HFI single-port PCIe-Gen3 ×16 in a 2:1 oversubscribed fat-tree topology.

Software. NNSA's Tri-Lab Operating System Stack (TOSS) version 3.4-4, which is based on RHEL 7.6 and Linux kernel 3.10.0-957.5.1. Key software components are Charliecloud 0.9.10, Shifter 18.03.0, Singularity 3.3.0-rc1, HPCG 3.0, STREAM 5.10, SysBench 1.0.17, OpenMPI 3.1.4, and GCC 4.8.5. We built one container image with Docker and converted it to each technology using its native tools.