A recent paper [179] describes the set of applications used at the National Energy Research Scientific

Computing Center (NERSC) and presents the results of a comparative benchmark of EC2 and three

supercomputers. NERSC is located at Lawrence Berkeley National Laboratory and serves a diverse

community of scientists; it has some 3,000 researchers and involves 400 projects based on some 600

codes. Some of the codes used are:

Community AtmosphereMode (CAM), the atmospheric component of Community Climate System

Model (CCSM), is used for weather and climate modeling.12 The code developed at NCAR uses

two two-dimensional domain decompositions – one for the dynamics and the other for remapping.

The first is decomposed over latitude and vertical level; the second is decomposed over longitude/

latitude. The program is communication-intensive; on-node/processor data movement and

relatively long MPI13 messages that stress the interconnect point-to-point bandwidth are used to

move data between the two decompositions.

General Atomic and Molecular Electronic Structure System (GAMESS) is used for ab initio quantum

chemistry calculations. The code, developed by theGordon Research Group at theU.S.Department

of Energy’s Ames Lab at Iowa State University, has its own communication library, the Distributed

Data Interface (DDI), and is based on the same program multiple data (SPMD) execution

model. DDI presents the abstraction of a global shared memory with one-sided data transfers, even

on systems with physically distributed memory. On the cluster systems at NERSC the program uses

socket communication; on the Cray XT4 the DDI uses MPI and only one-half of the processors

compute, whereas the other half are data movers. The program is memory- and communication intensive.

Gyrokinetic14 (GTC) is a code for fusion research.15 It is a self-consistent, gyrokinetic tridimensional

particle-in-cell (PIC)16 code with a nonspectral Poisson solver. It uses a grid that

follows the field lines as they twist around a toroidal geometry representing a magnetically confined

toroidal fusion plasma. The version of GTC used at NERSC uses a fixed, one-dimensional

domain decomposition with 64 domains and 64MPI tasks. Communication is dominated by nearestneighbor

exchanges that are bandwidth-bound. The most computationally intensive parts of GTC

involve gather/deposition of charge on the grid and particle “push” steps. The code is memory intensive

because the charge deposition uses indirect addressing.

IntegratedMapandParticle Accelerator Tracking Time (IMPACT-T) is a code for the prediction and

performance enhancement of accelerators. It models the arbitrary overlap of fields from beamline

elements and uses a parallel, relativistic PIC method with a spectral integratedGreen function solver.

This object-oriented Fortran90 code uses a two-dimensional domain decomposition in the y−z

directions and dynamic load balancing based on the domains. Hockney’s Fast Fourier Transform

(FFT) algorithm is used to solve Poisson’s equation with open boundary conditions. The code is

sensitive to the memory bandwidth and MPI collective performance.

MAESTRO is a low Mach number hydrodynamics code for simulating astrophysical flows.17 Its

integration scheme is embedded in an adaptive mesh refinement algorithm based on a hierarchical

system of rectangular, nonoverlapping grid patches at multiple levels with different resolutions;

it uses a multigrid solver. Parallelization is via a tridimensional domain decomposition using a

coarse-grained distribution strategy to balance the load and minimize communication costs. The

communication topology tends to stress simple topology interconnects. The code has a very low

computational intensity, it stresses memory latency, and the implicit solver stresses global communications.

The message sizes range from short to relatively moderate.

MIMD Lattice Computation (MILC) is a Quantum Chromo Dynamics (QCD) code used to study

“strong” interactions binding quarks into protons and neutrons and holding them together in the

nucleus.18 The algorithm discretizes the space and evaluates field variables on sites and links of a

regular hypercube lattice in four-dimensional space-time. The integration of an equation of motion

for hundreds or thousands of time steps requires inverting a large, sparse matrix. The Conjugate

Gradient (CG)method is used to solve a sparse, nearly singular matrix problem.Many CG iteration

steps are required for convergence; the inversion translates into tridimensional complex matrix vector

multiplications. Each multiplication requires a dot product of three pairs of tridimensional

complex vectors; a dot product consists of five multiply/add operations and one multiply. The

MIMD computational model is based on a four-dimensional domain decomposition. Each task

exchanges data with its eight nearest neighbors and is involved in the all-reduce calls with very

small payload as part of the CG algorithm. The algorithm requires gather operations from widely

separated locations in memory. The code is highly memory- and computational-intensive and it is

heavily dependent on prefetching.

PARAllel Total Energy Code (PARATEC) is a quantum mechanics code that performs ab initio

total energy calculations using pseudo-potentials, a plane wave basis set, and an all-band (unconstrained)

Conjugate Gradient (CG) approach. Parallel three-dimensional FFTs transform the wave

functions between real and Fourier space. The FFT dominates the run-time; the code uses MPI

and is communication-intensive. The code uses mostly point-to-point short messages. The code

parallelizes over grid points, thereby achieving a fine-grain level of parallelism. The BLAS3 and

one-dimensional FFT use optimized libraries (e.g., Intel’s MKL or AMD’s ACML), which results

in high cache reuse and a high percentage of per-processor peak performance.

The authors of [179] use the High-Performance Computing Challenge (HPCC) benchmark to compare

the performance of EC2 with the performance of three large systems at NERSC. HPCC19 is a

suite of seven synthetic benchmarks: three targeted synthetic benchmarks that quantify basic system

parameters that characterize individually the computation and communication performance and four

complex synthetic benchmarks that combine computation and communication and can be considered

simple proxy applications. These benchmarks are:

• DGEMM.20 The benchmark measures the floating-point performance of a processor/core. The memory

bandwidth does little to affect the results, since the code is cache-friendly. Thus, the results of

the benchmark are close to the theoretical peak performance of the processor.

• STREAM.21 The benchmark measures the memory bandwidth.

• The network latency benchmark.

• The network bandwidth benchmark.

• HPL.22 A software package that solves a (random) dense linear system in double precision arithmetic

on distributed-memory computers. It is a portable and freely available implementation of the High-

Performance Computing Linpack Benchmark.

• FFTE. Measures the floating-point rate of execution of double precision complex one-dimensional

Discrete Fourier Transform (DFT).

• PTRANS. Parallel matrix transpose exercises the communications whereby pairs of processors

communicate with each other simultaneously. It is a useful test of the total communications capacity

of the network.

• RandomAccess. Measures the rate of integer random updates of memory (GUPS).

The systems used for the comparison with cloud computing are:

Carver. A 400-node IBM iDataPlex cluster with quad-core Intel Nehalem processors running at

2.67 GHz and with 24 GB of RAM (3 GB/core). Each node has two sockets; a single Quad Data

Rate (QDR) IB link connects each node to a network that is locally a fat tree with a global two dimensional

mesh. The codes were compiled with the Portland Group suite version 10.0 and Open

MPI version 1.4.1.

Franklin. A 9,660-node Cray XT4; each node has a single quad-core 2.3 GHz AMD Opteron

Budapest processor with 8 GB of RAM (2 GB/core). Each processor is connected through a

6.4 GB/s bidirectional HyperTransport interface to the interconnect via a Cray SeaStar-2 ASIC.

The SeaStar routing chips are interconnected in a tridimensional torus topology in which each

node has a direct link to its six nearest neighbors. Codes were compiled with the Pathscale or the

Portland Group suite version 9.0.4.

Lawrencium. A 198-node (1,584 core) Linux cluster; a compute node is a Dell Poweredge 1950

server with two Intel Xeon quad-core 64-bit, 2.66-GHz Harpertown processors with 16GBofRAM

(2 GB/core). A compute node is connected to a Dual Data Rate InfiniBand network configured as a

fat tree with a 3:1 blocking factor. Codes were compiled using Intel 10.0.018 and Open MPI 1.3.3.

The virtual cluster at Amazon had four EC2 Compute Units (CUs), two virtual cores with two

CUs each, and 7.5 GB of memory (an m1.large instance in Amazon parlance). A Compute Unit

is approximately equivalent to a 1.0–1.2 GHz 2007 Opteron or 2007 Xeon processor. The nodes are

connected with gigabit Ethernet. The binaries were compiled on Lawrencium. The results reported in

[179] are summarized in Table 4.1.

The results in Table 4.1 give us some ideas about the characteristics of scientific applications likely

to run efficiently on the cloud. Communication-intensive applications will be affected by the increased

latency (more than 70 times larger then Carver) and lower bandwidth (more than 70 times smaller than

Carver).