OSS usage and benchmarking in HPC

Christian Külker

Debian Edu/ Skolelinux

2011-03-20

Contents

Supercomputer, Supercomputing and High Performance Computing

Supercomputer, Supercomputing and HPC

- What is a supercomputer and why does it matter?
- In general, a supercomputer is defined as a computer that is at the forefront of current processing capability, particularly with respect to the speed of calculation.
- Top 500 (http://www.top500.org)

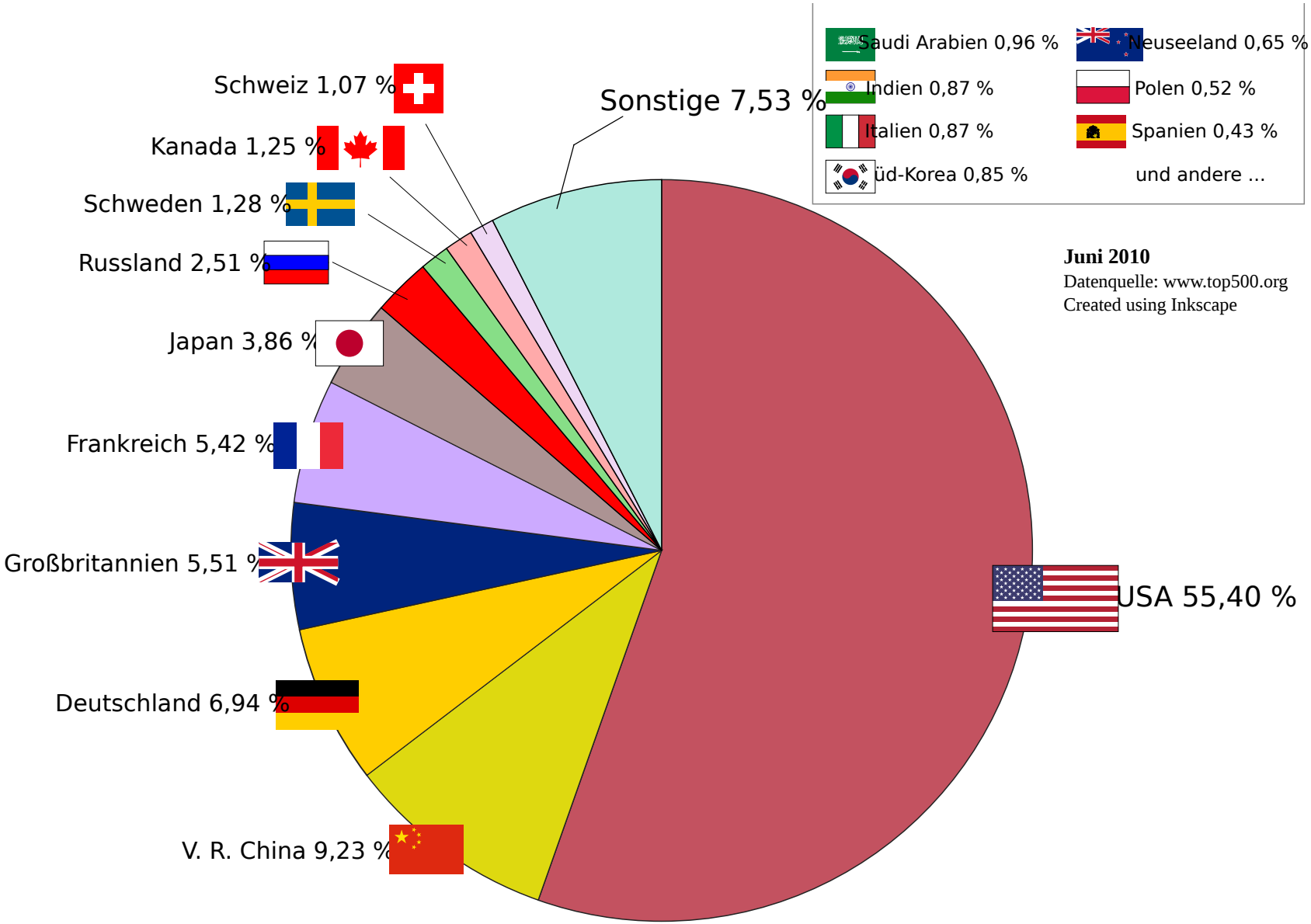
Supercomputing goes Green

- Green 500 (http://www.green500.org)
- In the context of The Green500 List, a supercomputer is a computing system that is fast enough to appear of the latest Top500 List.
- In the context of The Little Green500 List, a supercomputer is a computing system that achieves performance on the HPL benchmark at a high-enough level to have secured entry into the oldest Top500 list released within 19 months.
- http://www.green500.org/docs/pubs/RunRules_Ver0.9.pdf

The hit parade - Top 10 (ot of Top500)

1	Tianhe-1A	NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C
2	Jaguar	Cray XT5-HE Opteron 6-core 2.6 GHz
3	Nebulae	Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU
4	TSUBAME 2.0	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows
5	Hopper	Cray XE6 12-core 2.1 GHz
6	Tera-100	Bull bullx super-node S6010/S6030
7	Roadrunner	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband
8	Kraken XT5	Cray XT5-HE Opteron 6-core 2.6 GHz
9	JUGENE	Blue Gene/P Solution
10	Cielo	Cray XE6 8-core 2.4 GHz

Super Computing is a national sport



Gesamte Rechenleistung der Supercomputer in TOP500 pro Nation

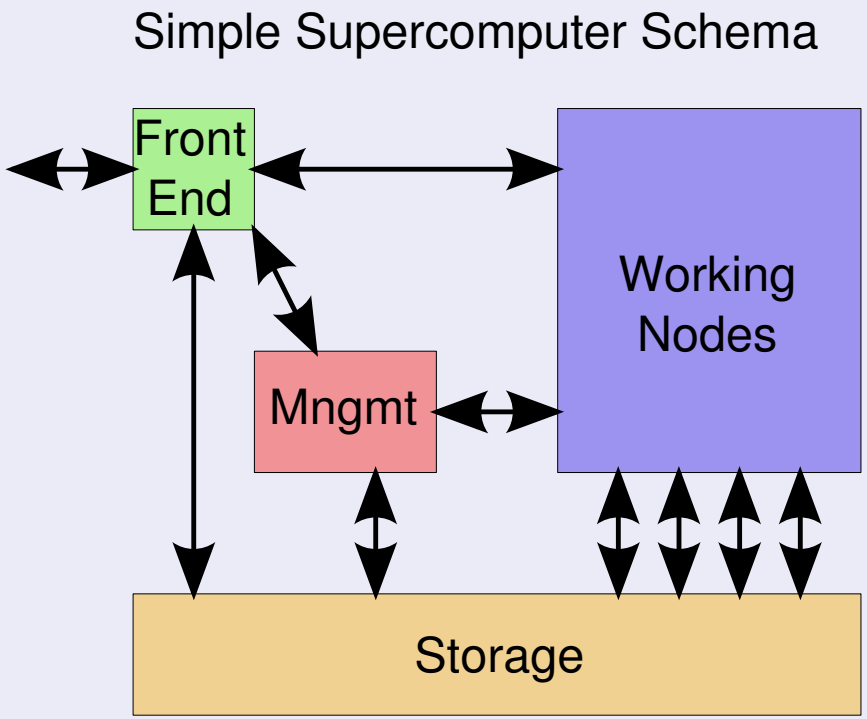
Different types of Supercomputers

- Scalar Processors <70th
- Vector processors >70th - mid 80th
- Parallel processing mid 80th - 90th
- Custom made processors (APE) and commodity processors (Intel,AMD, Alpha,...)
- modern supercomputers highly-tuned computer clusters using commodity processors combined with custom interconnects
- CPU/GPU and other accelerators (FPGA, ...)
- Different coupling (Strong: ApeMille, Loose: Cluster)
- Different Networks (Ethernet, Infiniband, Torus, Mesh, ...)

(Free?) Open Source Software usage

Simplified System Map

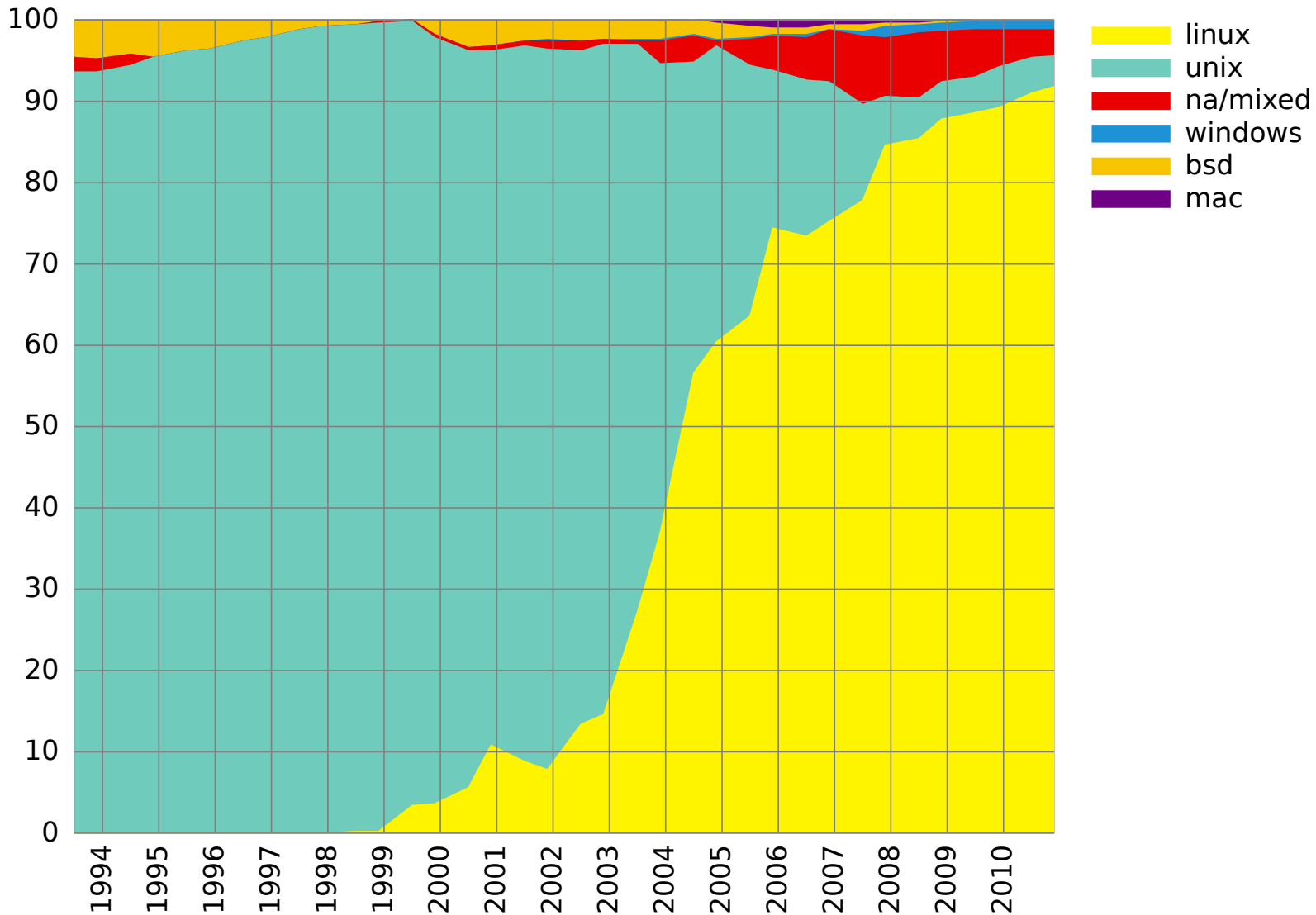
- Front End Servers
- Management Servers
- Storage /Common File System
- Working Nodes



Supercomputer Software Stack

- Management (User, Software, Node OS, Common File System, Resources)
- Sensor Network (Drivers, Aggregation, Event Triggering)
- User Application (Libraries, Compiler, Debuggers, Profilers, MPI, Environment, Special Purpose Libraries, Scheduling)
- Common File System (Monitoring, Mounting, Checking, Updating, Backup)
- Scheduler/ Queue (Workload, Accounting, Resources)
- System Analysis (Health, Utilisation, Accounting)

Operating systems used in Top500



Operating system usage in HPC

410	Linux
37	SuSE Linux (plus additions)
9	RedHat Linux (EHEL, and additions)
8	CentOS
36	Other OS

http://www.top500.org/stats/list/36/os

Good Services

- Mounted File System: NFS
- Handing out IPs: DHCP
- Name Service: DNS and BIND
- User Authentication: LDAP
- System Management: ?

The Debian View on HPC

- Cluster Provisioning: clobber, onesis, systemimage, FAI
- Cluster Management: cfengine, freeipmi, FAI, Kickstart, Autoyast/Alice
- Cluster Monitoring: Nagios, Ganglia, cacti
- Scheduler: slurm, SGE, Condor, TORQUE, MAUI
- Kernel Patches: (...)
- Userspace (mpich, openmpi, openfabric, Globus)

http://wiki.debian.org/HighPerformanceComputing

Open Source and Commercial Libraries

- NAMD: Molecular Dynamics
- VMD: Visual Molecular Dynamics
- mpiBLAST: Nucleotide/Protein Searching
- ... (a lot more)

=> Problem: License

MPI comes in different flavors

- MPICH
- LAM/MPI
- Open MPI
- Intel MPI
- HP MPI
- Microsoft Messaging Passing Interface
- OpenMP
- FT-MPI
- LA-MPI
- PACX-MPI
- Adaptive MPI

MPICH

MPICH1 derivatives

- MPICH 1.2.6..13 for Myrinet

MPICH2 derivatives

- IBM (MPI BlueGene/L and BlueGene/P)
- Cray (MPI over RedStorm and XT3)
- SiCortex (MPI SiCortex)
- Microsoft (MPICH2-MS)
- Intel (MPICH2-Nemesis)
- NetEffect (MPICH2-iWARP)
- Qlogic (MPICH2-PSM)
- Myricom (MPICH2-MX)
- Ohio State Univ. (MVAPICH and MVAPICH2)
- Univ. of British Columbia (MPICH2/SCTP)

OpenMPI

Open MPI represents the merger between three well-known MPI implementations:

- FT-MPI from the University of Tennessee
- LA-MPI from Los Alamos National Laboratory
- LAM/MPI from Indiana University

MPI Communication pattern

- point2point
- Collective: broadcast
- one2all
- all2one
- all2all n+1 -> n-1

MPI Fortran example

```
c  Fortran example
   program hello
   include 'mpif.h'
   integer rank, size, ierror, tag, status(MPI_STATUS_SIZE)

   call MPI_INIT(ierror)
   call MPI_COMM_SIZE(MPI_COMM_WORLD, size, ierror)
   call MPI_COMM_RANK(MPI_COMM_WORLD, rank, ierror)
   print*, 'node', rank, ': Hello world'
   call MPI_FINALIZE(ierror)
   end
```

Compiling and running a Program

- Compiling and Linking MPI Programs
- Linux with Ethernet or Infiniband (OpenMPI):

```
mpicc hello.c -o hello
```

- Running MPI Programs
- for OpenMPI create machine file (hosts to be used)
- define your resources CPU, Cores
- load your environment (mpi-selector)
- Example run (Default network, for example infinband):

mpirun

```
ckuelker@hpc> mpirun -np 4 -machinefile machinefile.hpc hello
    Process 0 on host01 out of 4
    Process 1 on host01 out of 4
    Process 3 on host02 out of 4
    Process 2 on host02 out of 4
```

- run via scheduler/ queue

```
qsub -pe nc0 16 ~/bin/hello-world.sh
```

Benchmarking in HPC

Best Practice of System Setup

- Choice of the correct benchark system
- qualify benchmak system
- make stress test
- do benchmark

Chroma

Production Code as benchmark

- LQCD (http://www.usqcd.org/)

Linpack

- Linpack is used for top500.org
- library for numerical linear algebra
- was written in Fortran by Jack Dongarra, Jim Bunch, Cleve Moler, and Gilbert Stewart
- intended for use on supercomputers in the 1970s and early 1980s
- superseeded by LAPACK
- HPL used for top500

High Perfromance Linpack (HPL)

- http://www.netlib.org/benchmark/hpl/
- solves a (random) dense linear system in double precision (64 bits) arithmetic
- portable as well as freely available implementation of the High Performance Computing Linpack Benchmark.

High Performance Computing Challange - HPCC

- HPL - the Linpack TPP benchmark which measures the floating point rate of execution for solving a linear system of equations.
- DGEMM - measures the floating point rate of execution of double precision real matrix-matrix multiplication.
- STREAM - a simple synthetic benchmark program that measures sustainable memory bandwidth
- PTRANS (parallel matrix transpose) - useful test of the total communications capacity of the network.
- RandomAccess - measures the rate of integer random updates of memory (GUPS).
- FFT - measures the floating point rate of execution of double precision complex one-dimensional Discrete Fourier Transform (DFT).
- Communication bandwidth and latency

http://icl.cs.utk.edu/hpcc/

Links

- http://www.linuxhpc.org
- http://debianclusters.org/index.php/Main_Page

Thank you for listening

Christian Külker

HPC Project Manager
Partnership Program Coordinator
Eurotech - ETH Lab - Business Unit HPC

christian@skolelinux.de

http://www.cipux.org/