EUROTECH
Imagine. Build. Succeed.

Challenges for Free Open Source Software Applications on Linux Supercomputers

Christian Külker

ETH Lab / Eurotech

2013-05-31

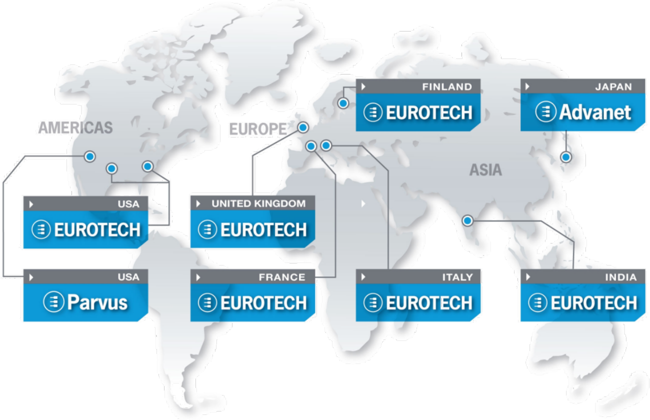
Eurotech Introduction

Eurotech is a listed global company (ETH.MI) that integrates hardware, software, services and expertise to deliver embedded computing platforms, sub-systems and high performance computing to leading OEMs, system integrators and enterprise customers for successful and efficient deployment of their products and services

The **Eurotech HPC division** aims to deliver advanced supercomputers and HPC solutions to enable science, technology and business to reach the excellence that will help the development of the humankind

Eurotech Introduction

Group Global Footprint



Eurotech Value Proposition

Products and Solutions for Core, Infrastructure, Edge

High Performance Computing Engines for the CORE	Connectivity Platforms to Build and Connect the EDGE	Components and Pervasive Devices for REAL WORLD Applications

Eurotech Introduction

Some of our Customers

THALES
NORTHROP GRUMMAN
Qadec
BOEING
GE Healthcare
Chevron
VARIAN medical systems
GENDEX
DRS TECHNOLOGIES
ALSTOM
GENERAL DYNAMICS
ZOE Medical
CNRS
L3 communications
BOMBARDIER
ACCRETECH
Nikon
DESY
HITACHI MEDICAL CORPORATION
BAE SYSTEMS
INFN Istituto Nazionale di Fisica Nucleare
Medtronic
LOCKHEED MARTIN
CUBIC CORPORATION
Johnson Controls
MITSUBISHI HEAVY INDUSTRIES, LTD.

HPC division highlights

- The Eurotech HPC division focuses on **designing, manufacturing, delivering and supporting high performance computing solutions**
- More than 14 years of history of delivering supercomputing systems and solutions to industry and academia
- First worldwide company to market hot water cooled high performance computers. First hot water cooled HPC in the market delivered in 2009.
- R&D capabilities nurtured in house and through collaboration with the best universities and research centres in Europe: INFN, Julich, Revensburg, Daisy…
- Funding member of ETP for HPC

Eurotech HPC project examples

Ape mille, 1999-2002
Ape next, 2002-2005

Janus, 2006-2008

Q-Pace, 2007-2009

Aurora Science, 2008-2010

Eurora 2012

Deep project 2012

Selex Elsag (E-security) 2011-2012

Supercomputing and High Performance Computing (HPC)

What is a supercomputer?

Supercomputer, Supercomputing and HPC

What is a supercomputer and why does it matter?

- A Supercomputer is a big computer
- »Super« stands for something extraordinary in terms of performance

> `https://en.wikipedia.org/wiki/Supercomputer`
>
> A supercomputer is a computer at the frontline of current processing capacity, particularly speed of calculation.

Supercomputer, Supercomputing and HPC

More practical approach ...

- An unambiguous definition do not exist, because the method of measuring the performance (speed of calculation) is not possible on all high performance computers in the same manner

http://www.top500.org/ http://www/green500.org

All computers out of the Top500 and Green500 list are Supercomputers.

- Often HPC is used instead of Supercomputing

How to measure performance? Example Top500

- Program: HPL 2.0 – High Performance Linpack
- Task: Performance number, measured in FLOPS
- FLOPS: Floating point operations per second
- Operation: Operation (multiplication) with numbers
- Floating point number: z.B. 1.528535047×10^5, or 152853.5047
- 1 PFLOPS = 1 PETA FLOPS = 1 000 000 000 000 000 FLOPS

Top 10 (of Top500.org) from November 2012 SLC

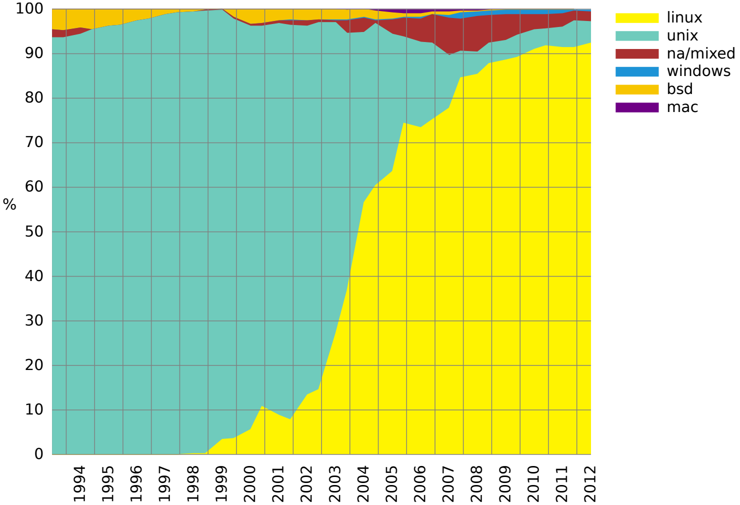
	Name	Computer	Site	OEM	Country	PFLOPS	OS
1	Titan	Cray XK7	DOE/SC/Oak Ridge National Laboratory	Cray Inc.	United States	17,590000	Linux
2	Sequoia	BlueGene/Q	DOE/NNSA/LLNL	IBM	United States	16,324751	Linux
3		K computer	RIKEN (AICS)	Fujitsu	Japan	10,510000	Linux
4	Mira	BlueGene/Q	DOE/SC/Argonne National Lab	IBM	United States	8,162376	Linux
5	JUQUEEN	BlueGene/Q	Forschungszentrum Juelich (FZJ)	IBM	Germany	4,141180	Linux
6	SuperMUC	iDataPlex DX360M4	Leibniz RZ	IBM	Germany	2897000	Linux
7	Stampede	PowerEdge C8220	Texas Adv. Comp. Center/Univ. of Texas	Dell	United States	2,660290	Linux
8	Tianhe-1A	NUDT YH MPP	National Supercomp. Center in Tianjin	NUDT	China	2,566000	Linux
9	Fermi	BlueGene/Q	CINECA	IBM	Italy	1,725492	Linux
10	DARPA Trial Subset	Power 775	IBM Development Engineering	IBM	United States	1,515000	Linux

What drives Supercomputers?

- The simple answer
- The complex answer

EUROTECH

Operating systems used in Top500 Nov. 2012



http://commons.wikimedia.org/wiki/File:Operating_systems_used_on_top_500_supercomputers.svg

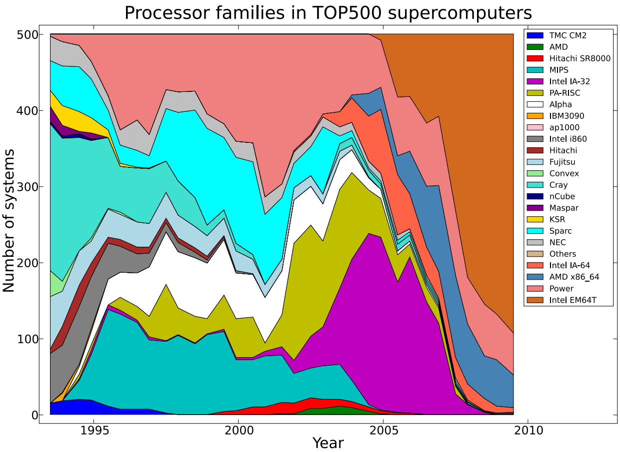
Author: Benedikt.Seidl; License: Public Domain;

Example Location of OS's inside a Supercomputer

- Login Node OS
- Head Node OS
- Chassis Controller Card OS
- Node Card
  - CPU OS
  - Board Management Controller OS
  - Accelerator OS (if any)
  - FPGA
- Maybe some other SOC OS (power supply, switches, ...)

Most of them are Linux. BMC, FPGA, SOC are separated

Processor Architectures used in Top500 - 2010



https://en.wikipedia.org/wiki/File:Processor_families_in_TOP500_supercomputers.svg

A Supercomputer Example Software Stack

EUROTECH

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: (...)
- User space (mpich, openmpi, openfabric, Globus)

http://wiki.debian.org/HighPerformanceComputing

Software

Aurora Software Integration services

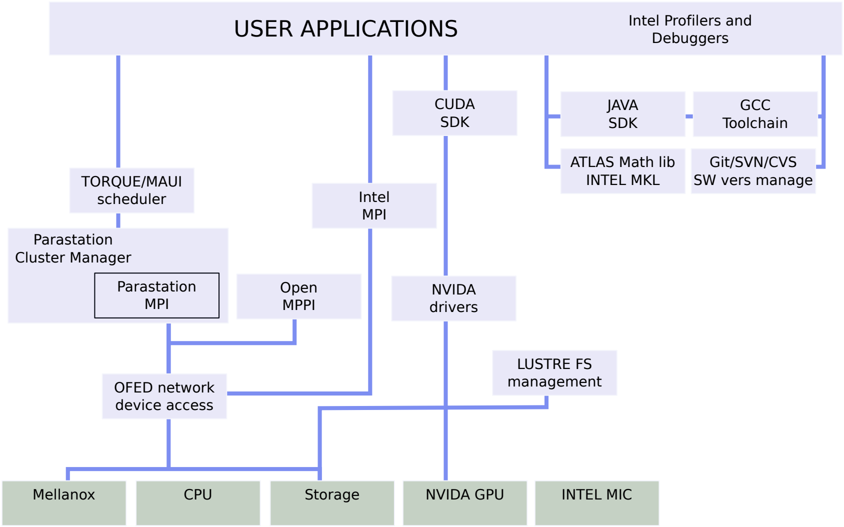
The Eurotech HPC team integrates drivers, operating systems, cluster managers, filesystems, resource managers, schedulers… to provide a turnkey environment where the customers run their applications.

Several applications can run on Eurotech Aurora supercomputers benefiting from a combination of compatibility and higher performance. For example

Abaqus , Avizo, ANSYS Workbench, ANSYS Fluent, Altair HyperGraph, Altair HyperMesh, Altair HyperView , Blender, EnSight.

Software

Software stack example



Supercomputer Applications

- User = scientist: parallel applications
- User = admin: Ganglia, Nagios, fs, ...
- User = system integrator: scheduler, ...
- User = OEM: BMC, Chassis Controller OS, FPGA, ...

Different interests in software: Not all interest satisfied with FOSS

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)

Challenges and Threats

EUROTECH

Challenges - The Supercomputer Paradigm

* Continuity is not a target, performance is

- frequent hardware changes
- frequent software changes
- frequent architecture changes
- frequent porting necessity

Challenge: Academia

Academic user, the forefront on high performance computers.

- We have basically free and open minded academia environment
- Nevertheless many of them have no deeper understanding of the concept and advantage of open source
- The one who publish (a paper) first will win (strange enough the code to write the paper is often not published) (the machines used are not available by the public)
- What counts is speed not openness. Proprietary or not does not matter
- Public funded. But no public demand for publish source code (at best: demand for paper)
- Academic code share under the impression to be a model of code share for favor.
- Code writing ability a necessity not a target as such (main discipline is not software: science, physics, ...

Challenge of a complex System

- Problem of using schedulers, and other blind tools for users.
- Non interoperability between super computers module system, miexec – mpirun (openmpi, intelmpi, ...)
- Different programming models THREAD, MPI, TREAD + MPI, OPENMP, OPENMP + MPI, ... (different MPI versions)
- No easy graphic output
- Different schedulers (pro on non pro versions)
- New type of CPU, accelerators makes porting a daily task, not an exception
- New Accelerators require new strategies (Custom -> FPGA -> Cell -> CPU -> GPU -> MIC)
- Companies with hybrid protection mechanism
- Not invented here syndrome

Threats for FOSS on Supercomputers 1

- BMC vendors (SOC) (No user applications, closed eco system)
- Too many hardware standards (I2C, GPIO, SMB BUS, ...)
- Hardware is not to be designed to work easy (cheap?)
- Closed Hardware firmware
- Not much support for Linux from Hardware vendors
- Compiler/CPU/GPU/Accelerator vendor " lock-in" , example: Intel, NVIDIA
- Many hardware vendors, who provide software, have no clue about GPL
- Custom hardware, not purchasable by mortal humans
- Hardware is often secret
- HPC can only be designed in a few regions on this planet
- Complete change of software stack with each architecture change

Threats for FOSS on Supercomputers 2

- Part of the success of a HPC is the software: custom improved software (software thought as of a secret weapon) Information about hardware or software only under NDA
- Open Source programmers fork too much
- Open Source programmers have no access to SC/BMC,SOC
- Public funding not aware of FOSS (software and papers are private assets of researcher)
- Technical complexity rises, private programmers need more work to get started
- Reliability: who is in charge if there is a kernel bug?
- Bugs in software can be embarrassing for scientists
- FOSS publication not a high priority for scientists

Threats for FOSS on Supercomputers in Japan

- Sys admin culture seldom
- No HPC support from Intel Japan; support via US/ Europe
- Rather Unix then Linux history of sys admins
- Pragmatism: OS do not matter as long as it works
- Culture of 便利 (benri)
- Orientation towards graphical systems (Windows, Mac)
- Vendor dependency, relationship

Dream of the Future

- All devices are available with open and commercial software stack: BMC, BIOS, Accelerator, ...
- Compile once run everywhere (in high performance mode)
- Error recovery - when one node die, an other take over the work
- FOSS uses all cores
- FOSS uses all accelerators, if possible
- FOSS uses MPI or other communication if possible to run in parallel
- Linux supports parallelism per default

Summary

EUROTECH

Summary

- Parallel apps on the rise, programming and usage is too difficult. Too many middle ware.
- App/ network and OS not failure tolerant
- FOSS lacks key support for HPC (app, compiler, lib)
- Supercomputers are closed resources
- Supercomputers are not binary compatible (at least env)
- Public funded HPC software development not forced to be FOSS
- Customer not aware about advantages, afraid of reliability
- Understanding of FOSS and FOSS licenses difficult
- Often scientists believe to have different motivations
- Users specialized in their subject not software engineers
- Scalability problems with some FOSS: file systems, schedulers

Thank you for listening

Christian Külker

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

c.kuelker@ethlab.com

http://www.ethlab.com/

http://christian.kuelker.info/speech/

Special thanks to: Götz Waschk (DESY) and many others.