VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean

HPC Infrastructure at IICT-BAS and installed software

Emanouil Atanassov
Institute of Information and Communication Technologies - BAS

Outline

- Hardware available at IICT-BAS
- Initial software
- Additional software
- Directions for future deployments

Bulgarian HPC resources

-SEEM

AVITOHOL at IICT-BAS

150x HP ProLiant SL250s Gen8 each with
2x Intel Xeon E5-2650 v2 (8C/16T),
64 GB DDR3-1866 RAM and
2x Intel Xeopn Phi 7120P
6x HP ProLiant DL380p Gen8 nodes with
2x Intel Xeon E5-2650v2 (8C/16T),
64 GB DDR3-1866 RAM
Infiniband 56 Gb/s FDR
Storage system with 96 TB

AVITOHOL

Total Performance:
RPeak:412.3 TFlop/s
RMax: 264.2 TFlop/s
Top 500 position: 389

HPCG cluster at IICT-BAS

36 blades BL 280c(2x Intel X5560(4C/8T); 24GB DDR3);
8 management nodes HP DL 380 G6(2x Intel X5560(4C/8T); 32GB DDR3);
2 HP ProLiant SL390s G7(2x Intel E5649(6C/12T);96GB DDR3)
8x nVidia TESLA M2090 per server;
2 HP SL270s Gen8 (2x Intel Xeon E5-2650 v2(8C/16T); 128GB DDR3)
Total number of Xeon Phi 5110P coprocessors: 9
Total 132TBs of system storage

TOTAL PERFORMANCE:
RPEAK:22.94 TFlop/s

NCSA IBM Blue Gene/P

8192 PowerPC 450 processors
4TBs of system memory
12TBs of system stotage
IBM proprietary interconnect with
2.5 μs latency and 10GBps bandwidth

TOTAL PERFORMANCE:
RPEAK:27.85 TFlop/s
RMAX:23.45 TFlop/s

PHYSON at Sofia University

53 Intel Xeon x86_64 processors
524Gibs of system memory
6.5TBs of system storage
2x nVidia Tesla M2090 graphics processors

TOTAL PERFORMANCE:
RPEAK:3.57 TFlop/s
RMAX: 3.22 TFlop/s

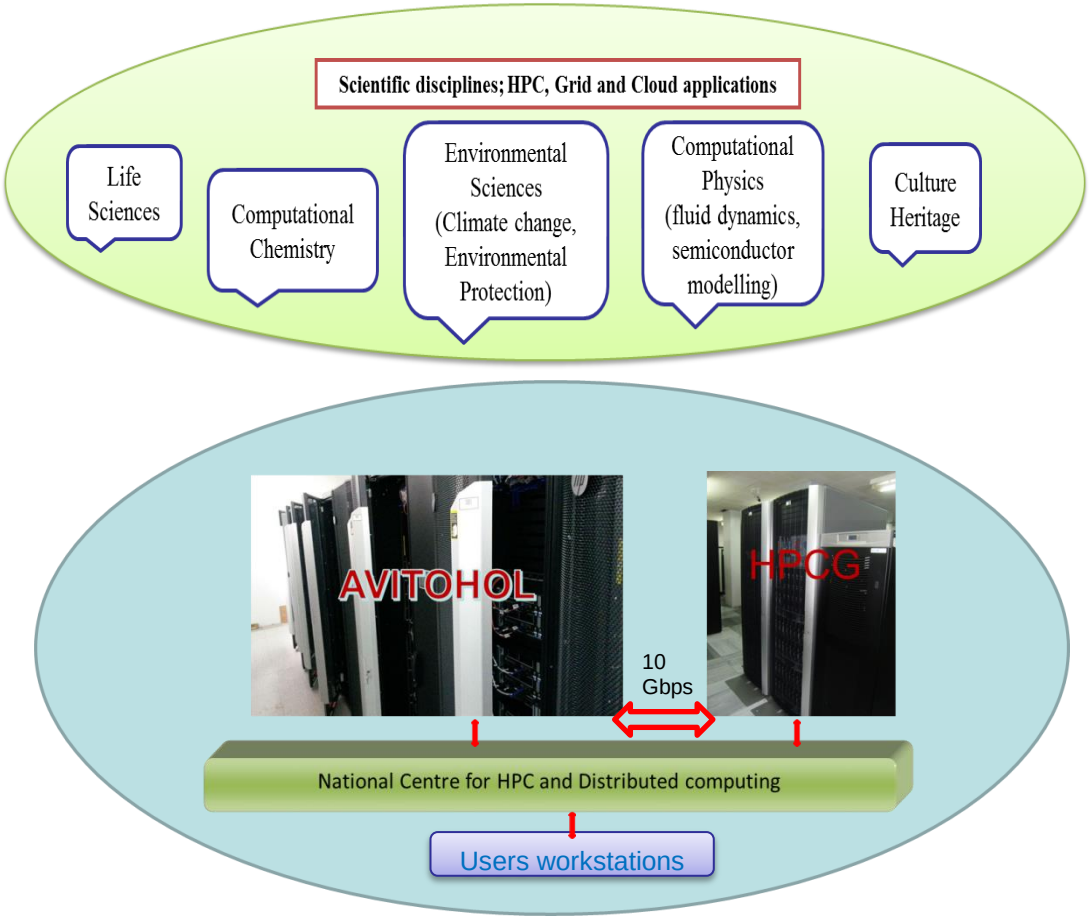
MADARA at IIOCCP-BAS

54 Primergy RX200 S5 servers with
2 Intel Xeon E5520(4C/8T) each
and a total of 800GB DDR3 1066MHz
20Gb/s DDR Infiniband
108TB System Storage by Fujitsu FibreCat SX100

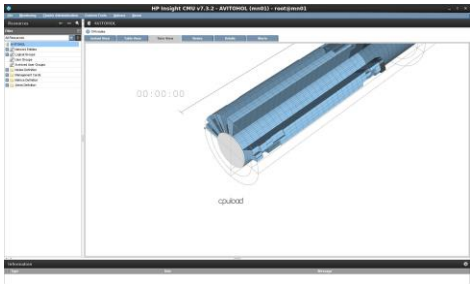
HPC&DC Center at IICT

Top500 List on 388th place (Nov 2015)
http://www.top500.org/system/178609

150 HP Cluster Platform SL250S GEN8 servers with 2 Intel Xeon E 2650 v2 CPUs and 2 Intel Xeon Phi 7120P coprocessors	
Site	IICT-BAS/Avitohol
Manufacturer	Hewlett-Packard
Cores	20700
Interconnection	FDR InfiniBand
Theoretical Peak Performance	412.3 Tflop/s
RMAX Performance	264.0 Tflop/s
Memory	9600 GB
Operation System	Red Hat Enterprise Linux for HPC
Storage capacity	96 TB SAN



Supercomputer System Avitohol - overview



- 150 dual-socket nodes HP ProLiant SL250S Gen8
- 4 I/O nodes HP ProLiant DL380p Gen8 with 2 Intel Xeon CPU E5-2650 v2, 64GB RAM, Fibre Channel cards
- 2 management nodes HP ProLiant DL380p Gen8 with 2 Intel Xeon CPU E5-2650 v2, 64GB RAM
- 96 TB of online disk space, connected with Fibre Channel with the I/O nodes.
- Fully non-blocking 56Gbps FDR InfiniBand network interconnecting all nodes
- The system consists of 8 water-cooled racks of type HP MCS 200, paired in couples.
- Each pair provides power and cooling for up to 50 kW of equipment, cooled by water cooling.
- About 90% of the computational power comes from the accelerators – one 7120P coprocessor achieves 1.25 TFlop/s in double precision, 2.5 TFlop/s in single precision

Supercomputer System Avitohol – compute nodes

Configuration of the computing nodes:

- Processors: dual Intel Xeon 8-core CPU E5-2650 v2 @ 2.6 GHz, hyperthreading enabled.
- Coprocessors: two Intel Xeon Phi 7120P coprocessors, 16 GB RAM and 61 cores each one
- 4-way hyperthreading enabled for them.
- Main memory of the compute nodes: 64 GB each (9.6 TB in total)
- Memory of the accelerators: 16 GB each (4.8 TB in total)
- One active Infiniband card/slot per server – used by CPUs and coprocessors.
- 4 dual water-cooled racks

Supercomputer System Avitohol – data storage and processing

- Storage system with 96 TB in 24 disks of 4 TB capacity.
- Accessible via Fibre Channel from the 4 I/O nodes.
- They export the data to the compute nodes.
- Currently this is done with the Lustre filesystem, which is a parallel filesystem, optimized for processing large files.
- Creation and deletion of large amount of small files is to be avoided.
- Other interfaces to data may be provided if necessary.

Auxillary HPCG cluster at IICT-BAS

- 36 blade servers HP BL 280c, deployed in 3 HP Cluster Platform Express 7000 enclosures, each with 2 Xeon X5560 @ 2.80GHz, 24 GB RAM – 576 cores total with more than 3 Tflops peak performance; 8 dual-socket HP DL 380 G6 with dual Intel X5560 @ 2.8 Ghz, 32 GB RAM. Non-blocking Infiniband interconnection @ 20 Gbps, 92% efficiency on LINPACK
- Total disk storage more than **132 T**B from three disk systems, interconnected with Fibre Channel.
- 2 HP ProLiant SL390s G7 4U servers with 16 NVIDIA Tesla M2090 graphic cards (total **8192 GPU cores** with **10.64 TFlops** in double precision); HP SL270s Gen8 4U server with 8 Intel Xeon Phi 5110P Coprocessors (total **480 cores**, 1920 threads, **8.088 TFlops** of double-precision peak performance per server).
- Total peak theoretical peak performance **22.93 TFlops**

Software configuration of Avitohol

- Red Hat Enterprise Linux for HPC on the compute nodes, series 6.
- Compatible version (Scientific Linux) on the auxiliary cluster.
- Internally, the Xeon Phi coprocessors have their own Linux operating system that run from RAM. It is called MPSS – version 3.6 currently.
- The Xeon Phi system appears like slower running computer, so software development (compilation/linking) is not expected to be done on it.
- The /home directory is visible also from the Xeon Phi
- The I/O is slower if done from the coprocessor

Software configuration of Avitohol – Resource management

- The resource management means distribution of work (jobs) to servers.
- It is controlled by the combination of Torque and MOAB.
- People should use only full servers at once.
- If a user is using the server it is expected that he or she also have exclusive control of the associated two coprocessors.
- Logging on with ssh to servers that are not part of your own job is disabled.
- Users should try
  - not to interfere with other people's jobs
  - not to produce high load on the login node
  - not to cause ``swapping'' – any software that requires this should be discussed before running it and any jobs that start to swap should be cancelled.
  - Swapping on the coprocessors is not possible.

Development software

- Compilers:
  - The Intel Fortan (ifort) and C/C++ (icc, icpc) compilers are the default compilers on the HPC cluster Avitohol and are provided automatically at login (see the output of **module list** for details on the version).
- To compile and link MPI codes using Intel compilers, use the commands mpiifort, mpiicc or mpiicpc, respectively.
- The GNU compiler collection (gcc, g++, gfortran) is available as well. A default version comes with the operating system. Version 4.4.7 is installed on Avitohol system. More recent versions can be accessed via environment modules. To compile and link MPI codes using GNU compilers, use the commands mpicc, mpic++/mpicxx, mpif77 or mpif90, respectively.

Development software

- The compilation for the Xeon Phi is usually done at the host, i.e., using cross-compilation.
- To load the development environment for the Xeon Phi one can issue

  source /opt/mpss/3.6/environment-setup-k1om-mpss-linux
- Then the version of the compiler and other development tools can be accessed with something like:

  k1om-mpss-linux-gcc

  k1om-mpss-linux-g++

  k1om-mpss-linux-nm
- Sometimes it is necessary to use newer versions of gcc either on the CPU or on Xeon Phi. Some such versions can be provided if necessary.

Development software

- Newer versions of python are frequently required.
- Python 2.7 is available
- Java (currently 1.8.0_72) is available
- R (3.3.0)

- Basically any popular open source software for x86_64 can be installed on the servers upon request.

Development libraries

- Many popular libraries are installed. Except from those that come with the OS, we mention the following:
  - Intel MKL - by default, an MKL environment module is already loaded. The module set the environment variables MKL_HOME and MKLROOT to the installation directory of MKL. The Intel MKL Link Line Advisor is often useful to obtain information on how to link programs with MKL. For example, to link statically with the threaded version of MKL on Avitohol (Linux, Intel64) using standard 32 bit integers, something like these must be added to the Makefile when invoking the Intel compiler:

```
-Wl,--start-group
$(MKLROOT)/lib/intel64/libmkl_intel_lp64.a
$(MKLROOT)/lib/intel64/libmkl_intel_thread.a
$(MKLROOT)/lib/intel64/libmkl_core.a
-Wl,--end-group -lpthread -lm –qopenmp
```

(in one line). If compiling directly in bash, one must use ${MKLROOT} instead of $(MKLROOT).

Development libraries

- **NetCDF** – various versions have been deployed. Some of them can be linked to run on the coprocessors
- **FFTW - t**he so-called "Fastest Fourier transforms in the West/World" software is installed. Currently the version of FFTW 3 that comes with the OS is 3.2.3 and is available in the standard locations for development software.
- **PETSc -** a suite of data structures and routines for the scalable (MPI parallel) solution of scientific applications modeled by partial differential equations. Available as a module.
- **SLEPc - l**ibrary for the solution of large scale sparse eigenvalue problems on parallel computers. Available as a module.
- **GSL -** the GNU Scientific Library (GSL) is a numerical library for C and C++ programmers (the FGSL FORTRAN add-on interface is installed). GSL provides a wide range of mathematical routines

Development tools

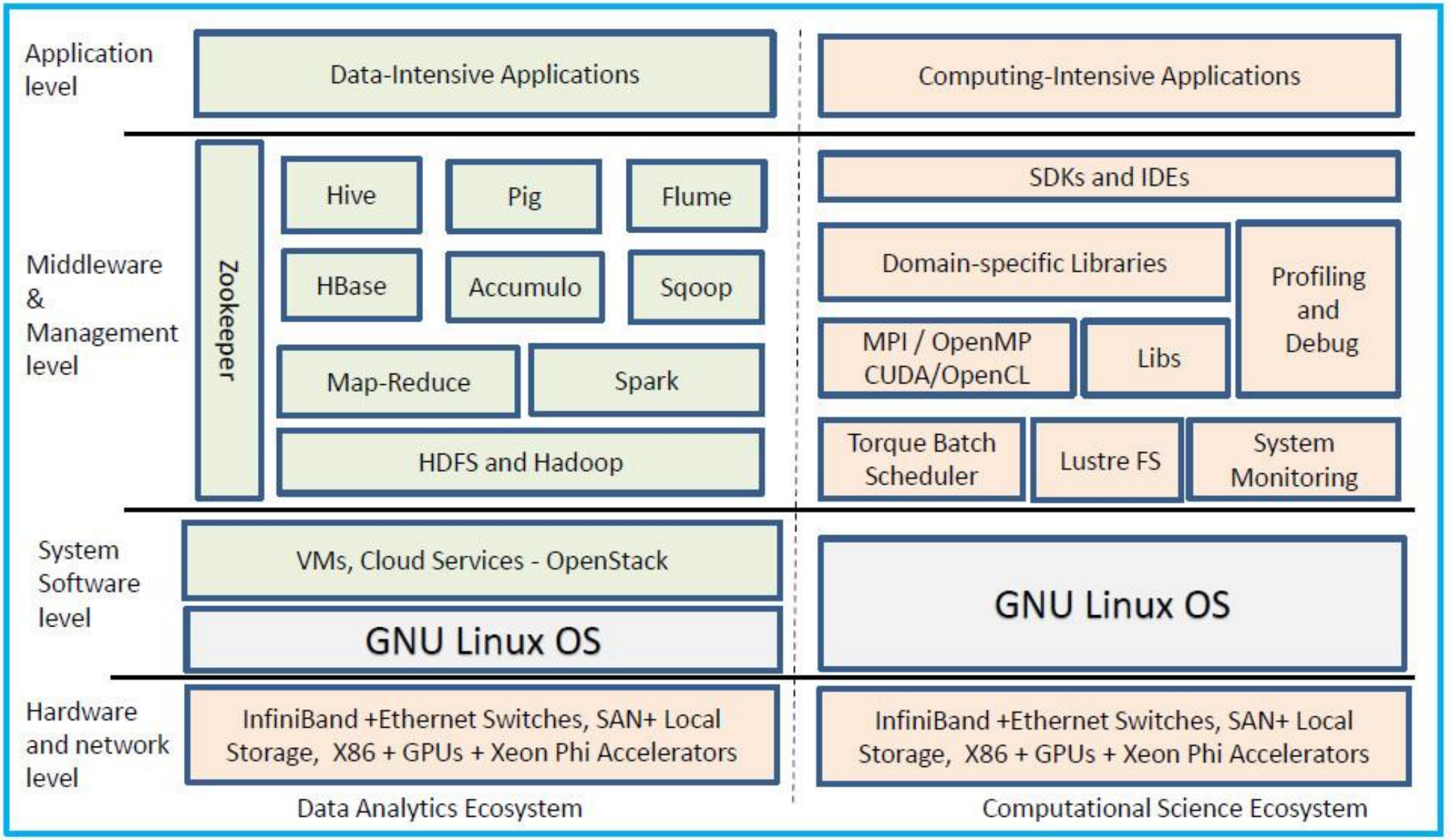
- **gdb**, the GNU debugger.
- Intel Inspector enables the debugging of threaded applications.
- If debugging features will not be used, the executable can be stripped of debugging information by running **strip.**
- **Intel VTune/Amplifier** is a powerful tool for analyzing the single core performance of a code – will be shown.
- **gprof**, the GNU profiler.
- **Intel Trace Analyzer and Collector** is a tool for profiling MPI communication.
- **Scalasca** enables the analysis of MPI/OpenMP/hybrid codes.
- Optimised executables can be obtained by first compiling for profile generation (option -prof-gen for Intel, --fprofile-generate for gcc) and then running the executable and using the generated profile with options like
  - -prof_use -prof_dir ./profdir for Intel
  - '-fprofile-use - for gcc, sometimes it is necessary to add -fprofile-correction or -flto-partition=none

Applications available

- SAP HANA – in-memory database useful for real-time data analysis.
- WRF and WRF-CHEM
- Gromacs
- GAMES
- CMAQ
- RegCM
- UKB

- Many others are in the process of being installed/tested.

Conclusions



Conclusions

- A powerful and versatile system is available now
- Considerable amount of software tools, libraries and applications have been deployed and tested.
- The system is flexible and can sustain large set of applications
- Optimal use of the system is achieved through multidisciplinary collaboration