Scientific Computing at MPP

Oliver Schulz

MAX-PLANCK-GESELLSCHAFT

Max-Planck-Institut für Physik
(Werner-Heisenberg-Institut)

oschulz@mpp.mpg.de

MPP Project Review, December 16, 2014

Outline

Computing Resources and Usage

(Selected) Software Projects

Data Preservation Efforts

Summary

Available Computing Resources

- In-house batch-system
- MPP Linux-cluster at RZG
- MPG supercomputer Hydra at RZG
- Experiment-specific resources (Grid, . . .)

In-House Batch-System

- Condor batch system, utilizes spare computing capacity on user workstations (Ubuntu and SUSE Linux)
- Computing capacity: 188 nodes, 1001 cores, 300 GB RAM
- Storage capacity: 70 (soon 130) TB total net space (CephFS, not available on all nodes yet)

Max-Planck-Institut für Physik
(Werner-Heisenberg-Institut)

In-House Batch-System

- Condor batch system, utilizes spare computing capacity on user workstations (Ubuntu and SUSE Linux)
- Computing capacity: 188 nodes, 1001 cores, 300 GB RAM
- Storage capacity: 70 (soon 130) TB total net space (CephFS, not available on all nodes yet)
- Currently mainly used by theory (low-IO applications)
- Soon: IO-intensive applications possible due to CephFS and increased network bandwidth

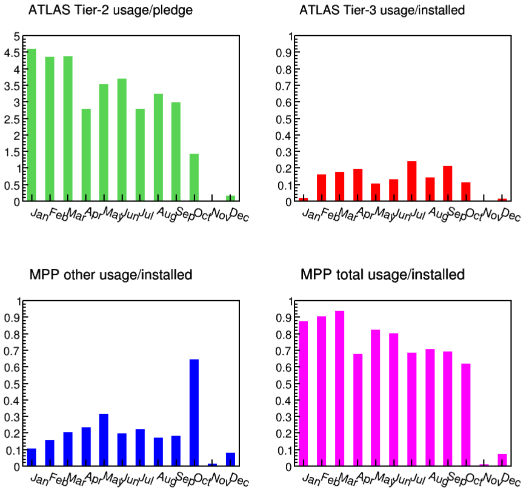
MPP Linux-Cluster at RZG

- Computing capacity: 160 nodes, 1776 cores, 3.5 TB RAM
- Storage capacity: 200 TB Storage (GPFS), 1.5 PB dCache
- Operating system: SLC-6

MPP Linux-Cluster at RZG

- Computing capacity: 160 nodes, 1776 cores, 3.5 TB RAM
- Storage capacity: 200 TB Storage (GPFS), 1.5 PB dCache
- Operating system: SLC-6
- Users: ATLAS Tier-2/3, MAGIC analysis centre, theory, GERDA, ILC, BELLE(II)
- Front-end nodes: mppui[1-3].t2.rzg.mpg.de

MPP Linux-Cluster Utilization 2014



ATLAS Tier-2 usage/pledge

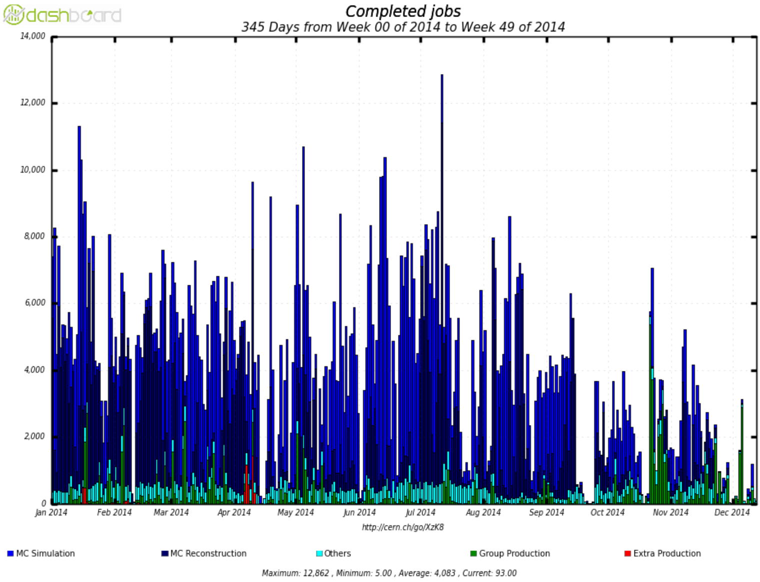
ATLAS Tier-3 usage/installed

MPP other usage/installed

MPP total usage/installed

Less ATLAS production at the Moment, but will ramp back up early 2015 (Run-2 MC production).

MPP Cluster ATLAS Jobs 2014



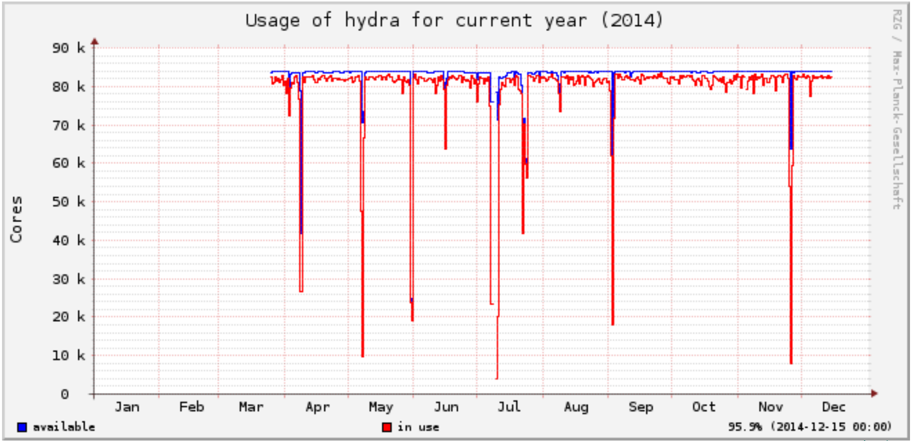
MPG supercomputer Hydra at RZG

- First stage (610 Sandy Bridge nodes) since Sept 2012, main part (3500 Ivy Bridge nodes) installed October 2013
- Total: 4110 nodes, 83000 cores, 280 TB RAM
- Storage capacity: 4.5 PB (GPFS, 0.75 PB perm., rest temp.)
- Peak performance 1.7 PetaFlop/s

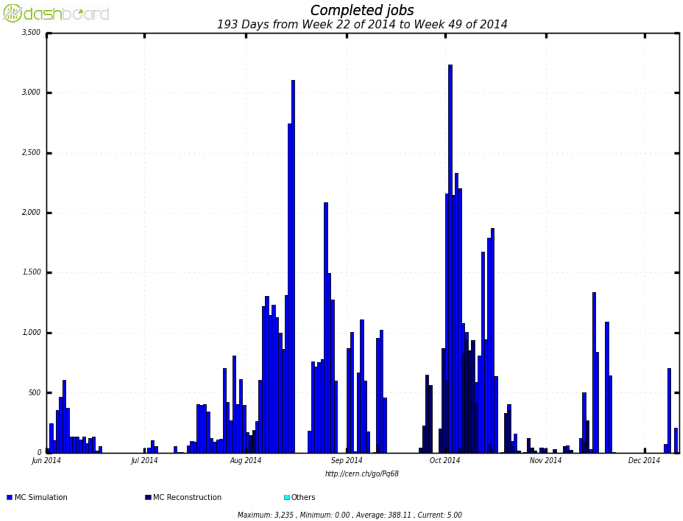
MPG supercomputer Hydra at RZG

- First stage (610 Sandy Bridge nodes) since Sept 2012, main part (3500 Ivy Bridge nodes) installed October 2013
- Total: 4110 nodes, 83000 cores, 280 TB RAM
- Storage capacity: 4.5 PB (GPFS, 0.75 PB perm., rest temp.)
- Peak performance 1.7 PetaFlop/s
- Fast InfiniBand FDR14 interconnect, 5 domains with internal fat-tree topology
- Contains 338 NVIDIA GPU nodes (1 PetaFlop/s total) and 12 Intel Xeon Phi nodes

Hydra Utilization 2014



ATLAS at Hydra 2014


Completed jobs
193 Days from Week 22 of 2014 to Week 49 of 2014

Maximum: 3,235 , Minimum: 0.00 , Average: 388.11 , Current: 5.00

- Grid-integration by Luca Mazzaferro via ARC-CE 4.1.0
- Currently limited to MC jobs due due to IO Limitations

New Archive System at RZG

- New tape-archive system at RZG installed in 2014
- Hierarchical storage management based on HPSS, data automatically moved between disk and tape
- Current capacity 7.5 PB, extensible

New Archive System at RZG

- New tape-archive system at RZG installed in 2014
- Hierarchical storage management based on HPSS, data automatically moved between disk and tape
- Current capacity 7.5 PB, extensible
- Access via host archive.rzg.mpg.de, personal archives at /ghi/r/<userid-initial>/<userid>
- Data transfer from login nodes or MPP cluster via scp / rsync / sshfs or similar
- Directly mounted on hydra, easy to move data between GPFS and archive

New Archive System at RZG

- New tape-archive system at RZG installed in 2014
- Hierarchical storage management based on HPSS, data automatically moved between disk and tape
- Current capacity 7.5 PB, extensible
- Access via host archive.rzg.mpg.de, personal archives at /ghi/r/<userid-initial>/<userid>
- Data transfer from login nodes or MPP cluster via scp / rsync / sshfs or similar
- Directly mounted on hydra, easy to move data between GPFS and archive
- Avoid small files - zip or tar up what you archive (aim for 1 GB to 500 GB file size)
- Archiving keeps your data safe (copy is stored at LRZ)
- Go easy on your colleagues and our budget: Move old data to archive!

Software Projects at MPP (and beyond)

- Efficient and reliable computing depends on high-quality software tools
- Avoid re-inventing the wheel, pool resources, release as open source

Software Projects at MPP (and beyond)

- Efficient and reliable computing depends on high-quality software tools
- Avoid re-inventing the wheel, pool resources, release as open source
- (Selected) success stories: CUBA, BAT, GoSam and SecDec
- New project: DatABriCxx

Cuba

Multidimensional numerical integration

- Motivation: Very common problem,
  but efficient and stable solutions highly non-trivial
- Developers: Thomas Hahn et al.
- Four different integration algorithms,
  all with C/C++, Fortran, and Mathematica interface

Cuba

Multidimensional numerical integration

- Motivation: Very common problem, but efficient and stable solutions highly non-trivial
- Developers: Thomas Hahn et al.
- Four different integration algorithms, all with C/C++, Fortran, and Mathematica interface
- Multi-purpose, used in many physics software projects
- Also non-physics / industry users
- New release 4.1 in November 2014

Cuba

Multidimensional numerical integration

- Motivation: Very common problem,
  but efficient and stable solutions highly non-trivial
- Developers: Thomas Hahn et al.
- Four different integration algorithms,
  all with C/C++, Fortran, and Mathematica interface
- Multi-purpose, used in many physics software projects
- Also non-physics / industry users
- New release 4.1 in November 2014
- Automatic parallelization: Supports vectorization,
  multi-core and GPU computing
- Homepage: http://www.feynarts.de/cuba/

BAT: Bayesian Analysis Toolkit

- Motivation: Bayes' theorem simple on paper, but numerics are hard
- Allen Caldwell et al. - currently 7 developers at MPP, TUM, Universe Cluster, TU-Dortmund
- Some prominent use cases:
  - ATLAS Z' search - Phys. Lett. B 719 (2013)
  - GERDA Phase-I Analysis - Phys. Rev. Lett. 111 (2013)
  - UTFIT: D meson mixing - arXiv:1402.1664
  - PAMELA: cosmic-ray proton spectrum - arXiv:1306.1354
- Optionally uses Cuba for integration
- New release 0.9.4 in November 2014, version 1.0 soon

BAT: Bayesian Analysis Toolkit

- Motivation: Bayes' theorem simple on paper, but numerics are hard
- Allen Caldwell et al. - currently 7 developers at MPP, TUM, Universe Cluster, TU-Dortmund
- Some prominent use cases:
  - ATLAS Z' search - Phys. Lett. B 719 (2013)
  - GERDA Phase-I Analysis - Phys. Rev. Lett. 111 (2013)
  - UTFIT: D meson mixing - arXiv:1402.1664
  - PAMELA: cosmic-ray proton spectrum - arXiv:1306.1354
- Optionally uses Cuba for integration
- New release 0.9.4 in November 2014, version 1.0 soon
- Started work on BAT-2: Re-design, parallel (multi-core and multi-node), more algorithms, C++11
- Homepage: https://www.mppmu.mpg.de/bat/

GoSam

Automated calculation of one-loop amplitudes
(for multi-particle processes in renormalizable QFT)

- GoSam collaboration, Gudrun Heinrich et al., 11 members
- Link to phenomenological analysis/experiment
- Interface to Monte-Carlo programs
- Easy to use
- Popular in the community

GoSam

Automated calculation of one-loop amplitudes
(for multi-particle processes in renormalizable QFT)

- GoSam collaboration, Gudrun Heinrich et al., 11 members
- Link to phenomenological analysis/experiment
- Interface to Monte-Carlo programs
- Easy to use
- Popular in the community
- Version 2.0 released 2014 (arXiv:1404.7096):
  Improved code generation, new reduction methods, extended applicability, easy installation
- Homepage: http://gosam.hepforge.org/

SecDec

Numerical evaluation of dimensionally regulated parameter integrals

- Motivation: How to find BSM physics without "smoking gun"? Precision calculations!
- Developers: G. Heinrich, S. Borowka, J. Carter
- Sector decomposition algorithm (T. Binoth, G. Heinrich)
- Languages: Mathematica, Fortran/C

SecDec

Numerical evaluation of dimensionally regulated parameter integrals

- Motivation: How to find BSM physics without "smoking gun"? Precision calculations!
- Developers: G. Heinrich, S. Borowka, J. Carter
- Sector decomposition algorithm (T. Binoth, G. Heinrich)
- Languages: Mathematica, Fortran/C
- Builds on Cuba
- Widely used in the community
- Version 2.1 released in 2014:
  Very useful for 2-loop problems with several mass scales
- Already running jobs on Hydra - theory needs HPC too!
- Homepage: http://secdec.hepforge.org/

Max-Planck-Institut für Physik
(Werner-Heisenberg-Institut)

DatABriCxx

Data analysis bricks in C++

- Motivation: Modular Analysis on multiple loop levels (runs, events, channels, . . .), easy code re-use
- Developers: Oliver Schulz / GeDet Group
- Developed for GeDet, may find it's way into GERDA, interest from CRESST and another (external) collaboration

DatABriCxx

Data analysis bricks in C++

- Motivation: Modular Analysis on multiple loop levels (runs, events, channels, . . .), easy code re-use
- Developers: Oliver Schulz / GeDet Group
- Developed for GeDet, may find it's way into GERDA, interest from CRESST and another (external) collaboration
- Variant of data-flow and map/reduce processing models: Bricks with input, output and parameter terminals, combination defined via config file
- Languages: C++11, JSON
- Based on ROOT-6, but also suitable for non-ROOT data

DatABriCxx

Data analysis bricks in C++

- Motivation: Modular Analysis on multiple loop levels (runs, events, channels, . . .), easy code re-use
- Developers: Oliver Schulz / GeDet Group
- Developed for GeDet, may find it's way into GERDA, interest from CRESST and another (external) collaboration
- Variant of data-flow and map/reduce processing models: Bricks with input, output and parameter terminals, combination defined via config file
- Languages: C++11, JSON
- Based on ROOT-6, but also suitable for non-ROOT data
- Ready for serious use in early 2015
- Interested parties welcome to join in early
- Homepage: https://github.com/mppmu/databricxx

Data Preservation

- Huge investment in past experiments (HERA, LEP, . . .)
- New discoveries (e.g. BSM physics at LHC) can make people go back to old data
- Preserve capability to run new analysis on old data
- DPHEP: ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics

Data Preservation

- Huge investment in past experiments (HERA, LEP, ...)
- New discoveries (e.g. BSM physics at LHC) can make people go back to old data
- Preserve capability to run new analysis on old data
- DPHEP: ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics
- Time is our friend: Data stays (mostly) constant, storage and processing becomes cheaper and cheaper
- Time is our enemy: Rapid loss of know-how after experiment ends and collaboration dies
- Also need solutions for smaller / non-collider experiments

DPHEP at MPP

DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

- Departments Bethke and Caldwell
- Gained a lot of momentum at MPP during last year (S. Kluth, A. Verbytskyi)
- Past experiments with MPP involvement:
  H1, ZEUS, JADE, OPAL
- Data stored at DESY and at RZG (dCache)

DPHEP at MPP

DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

- Departments Bethke and Caldwell
- Gained a lot of momentum at MPP during last year (S. Kluth, A. Verbytskyi)
- Past experiments with MPP involvement: H1, ZEUS, JADE, OPAL
- Data stored at DESY and at RZG (dCache)
- Huge effort by all four experiments to preserve electronic documentation: Complete
- Can't maintain code forever → virtualization
- Key elements: (Automatic) verification and examples

DPHEP at MPP

Study Group for Data Preservation and Long Term Analysis in High Energy Physics

- Departments Bethke and Caldwell
- Gained a lot of momentum at MPP during last year (S. Kluth, A. Verbytskyi)
- Past experiments with MPP involvement: H1, ZEUS, JADE, OPAL
- Data stored at DESY and at RZG (dCache)
- Huge effort by all four experiments to preserve electronic documentation: Complete
- Can't maintain code forever → virtualization
- Key elements: (Automatic) verification and examples
- Current experiments (e.g. LHC): Prepare for preservation early!

Example: ZEUS Data Preservation

- Collaboration has defined common n-tuple format,
carefully chosen for future analysis,
all calibration and corrections applied
- MC-Datasets for all relevant processes
(signals and backgrounds)

Example: ZEUS Data Preservation

- Collaboration has defined common n-tuple format, carefully chosen for future analysis, all calibration and corrections applied
- MC-Datasets for all relevant processes (signals and backgrounds)
- Virtual machine with all software:
  - Scientific Linux 7, 64 bit
  - Kickstart installation from custom ISO image
  - Contains all software: Compilers, ROOT, PAW, Event display (ZEVIS), file catalog (cninfo), stand-alone MC production package (ZMSP), . . .
  - Not tied to specific storage technology

Example: ZEUS Data Preservation

- Collaboration has defined common n-tuple format, carefully chosen for future analysis, all calibration and corrections applied
- MC-Datasets for all relevant processes (signals and backgrounds)
- Virtual machine with all software:
  - Scientific Linux 7, 64 bit
  - Kickstart installation from custom ISO image
  - Contains all software: Compilers, ROOT, PAW, Event display (ZEVIS), file catalog (cninfo), stand-alone MC production package (ZMSP), ...
  - Not tied to specific storage technology
- Electronics documentation collected and prepared (DESY): Web-pages decoupled from databases, PDFs, etc.

Smaller / Non-Collider Experiments

- Challenge: More projects, less resources per project
- Find common strategies for in-house projects (CRESST, GERDA, GeDet, CRESST, MAGIC, . . .)
- Strive for well structured, easy to use software now - makes preservation easier later

Smaller / Non-Collider Experiments

- Challenge: More projects, less resources per project
- Find common strategies for in-house projects (CRESST, GERDA, GeDet, CRESST, MAGIC, . . .)
- Strive for well structured, easy to use software now - makes preservation easier later
- In general: Build / use common computing resources for MPP data preservation efforts
- Explore promising new hardware-abstraction technologies (Docker, CoreOS, . . .)

Summary

- Substantial computing resources available at MPP and RZG - choose the right one for the job
- Additions this year:
  - New supercomputer Hydra
  - New archive system
  - Extended in-house storage (CephFS)

Summary

- Substantial computing resources available at MPP and RZG - choose the right one for the job
- Additions this year:
  - New supercomputer Hydra
  - New archive system
  - Extended in-house storage (CephFS)
- MPP very active in various software projects with high reputation and broad applicability

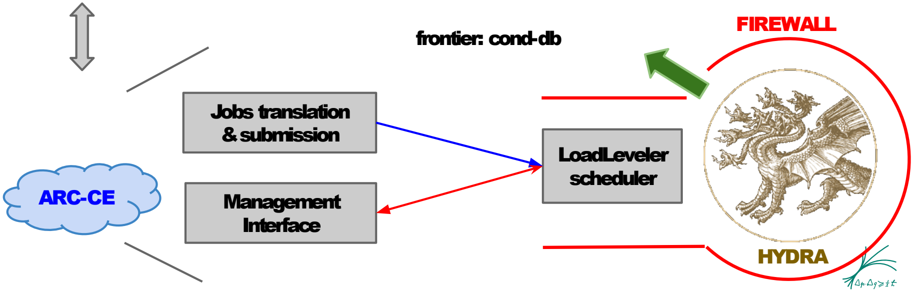
Summary

- Substantial computing resources available at MPP and RZG - choose the right one for the job
- Additions this year:
  - New supercomputer Hydra
  - New archive system
  - Extended in-house storage (CephFS)
- MPP very active in various software projects with high reputation and broad applicability
- Data preservation efforts well underway for past experiments - prepare early for the current ones

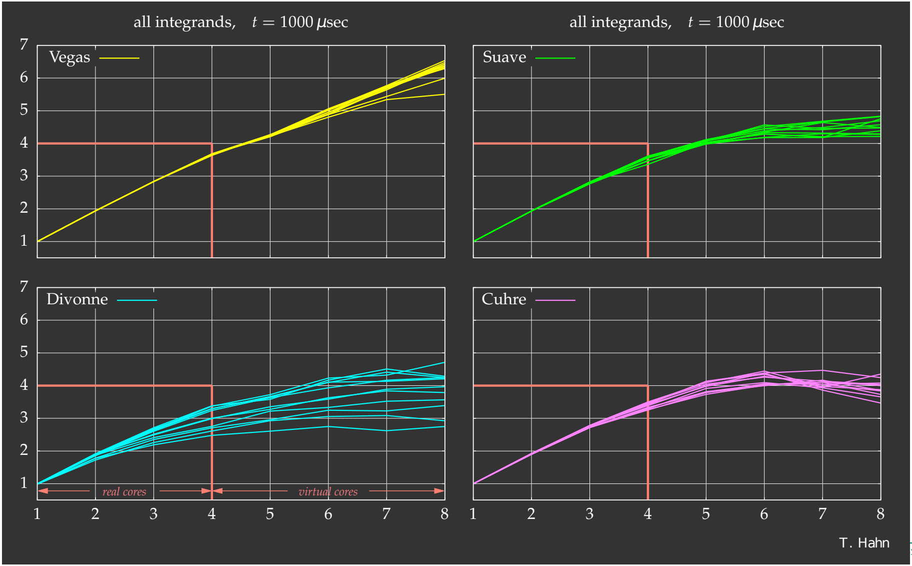
Appendix

HYDRA/ARC-CE architecture

1. **HYDRA system is accessible only from inside the MPG network.**
2. **ATLAS jobs have to be submitted via arcControlTower which interacts with ARC-CE.**
3. **ARC-CE "translates" the job description in the LoadLeveler "language" and submits the job.**
4. **ARC-CE takes also care of**
   - a. **monitoring the job status;**
   - b. **managing and storing the jobs results;**
   - c. **providing informations about jobs to the grid.**

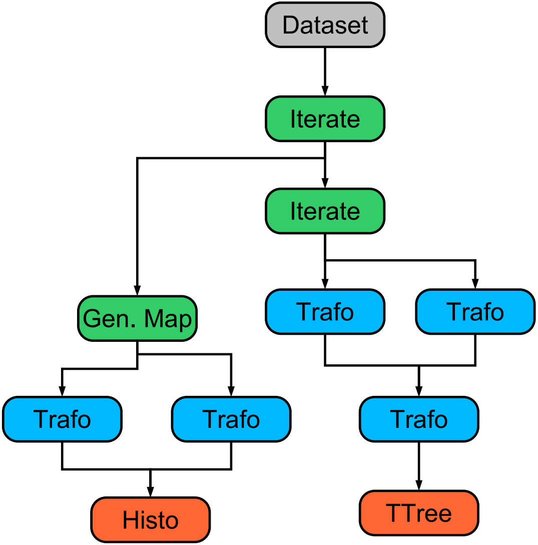


CUBA Multi-Core Performance



T. Hahn

DatABriCxx Data Flow



Brics form a directed, acyclic graph (DAG)