

Scientific computing and data preservation at MPP

MPP Project review 2020
Stefan Kluth
14.12.2020

People

- MPP IT Fachabteilung (T. Hahn)
 - Manuel Krämer, Konstantinos Kiriakidis, Alfred Kriesel, Katrin Krebs, Uwe Leupold, Yaozhi Pan
- MPP at MPCDF (SK)
 - Cesare Delle Fratte, Sergio Tafula
- MPP Computing Commission
 - I. Abt, S. Bethke, T. Hahn, SK (chair), D. Paneque, F. Simon, O. Schulz, S. Stonjek
 - Meetings (generally) public

Overview

Activity

O(1000 or more) batch jobs,
GPU support

O(100) Batch jobs

Programming, interactive cptg
E-mail, web, documents, etc

Compute systems

MPCDF systems: COBRA, DRACO

MPP condor batch jobs

MPP desktop PCs, few developer
Machines w/ GPU and large RAM

MPCDF

COBRA: 3424 compute nodes, 136,960 cores,
128 Tesla V100-32 GPUs, 240 Quadro RTX
5000 GPUs, 529 TB RAM, 7.9 TB HBM2,
since 2018

DRACO: >900 compute nodes, 30.688 cores,
212 GTX980 GPUs, 128 TB RAM,
since 2016



MPCDF MPP cluster

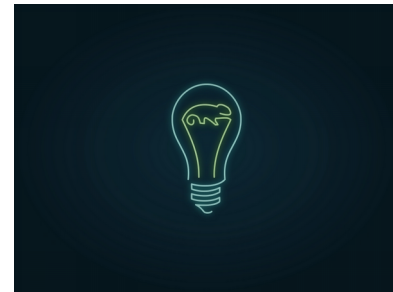
- 130 (+special) nodes, >3500 cores, 4 GB/core
 - Node groups at, bt, ft, kt, login nodes mppui1, 2, 3
- Large RAM nodes zt1, zt2 (Henn group):
 - Zt1: 6 TB, 192 cores, zt2: 3 TB, 36 cores
- More than 5 PB (soon) storage
 - /ptmp/mpp (gpfs), dCache
- CentOS7, Slurm batch, singularity, /cvmfs

MPP desktop PCs

- Theory groups
 - > 80 PCs \geq 8 core, 2-4 GB/core, ssd
 - opensuse tumbleweed, condor, /remote/ceph, /cvmfs
- Expt. groups
 - > 100 PCs \geq 8 core, 2-4 GB/core, ssd
 - Ubuntu 18 LTS, condor, /remote/ceph, /cvmfs
- Common
 - CEPH storage > 2 PB
 - Few servers with 512-1024 GB RAM, Nvidia GPUs, R&D, local jobs

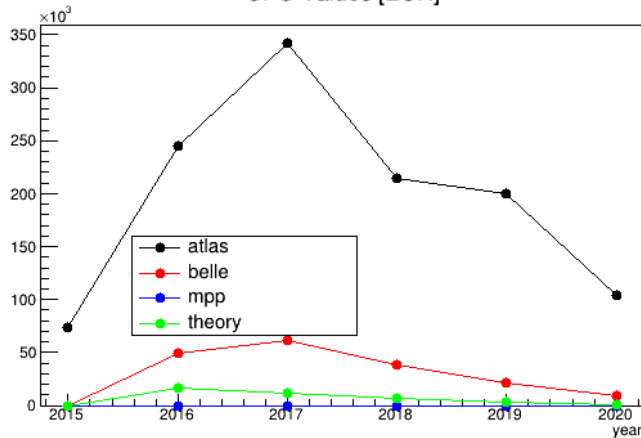
MPP 1 Linux

- Plan to move all desktop PCs to common (one) Linux
 - Rolling release (opensuse tumbleweed)
 - Desktops, Singularity, /cvmfs, /remote/ceph, condor
- Test machines available
 - Please test your (local) workflows
- Rolling release Linux
 - Guarantee support of modern hardware
 - Decent and recent desktop software
 - Scientific workflows in containers, data in /remote/ceph, batch jobs with condor

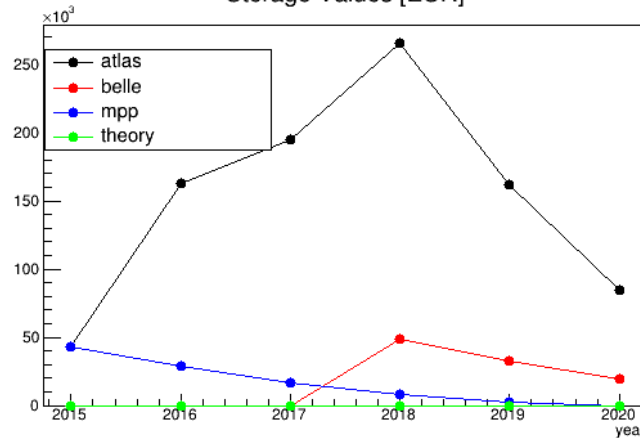


MPCDF MPP cluster

CPU Values [EUR]



Storage Values [EUR]



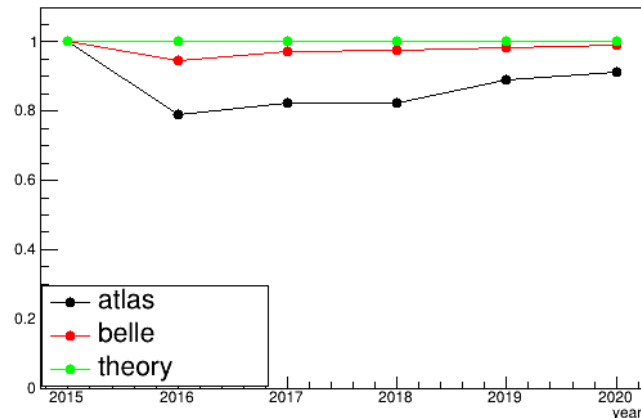
Status Sep 2020

New procurements coming:

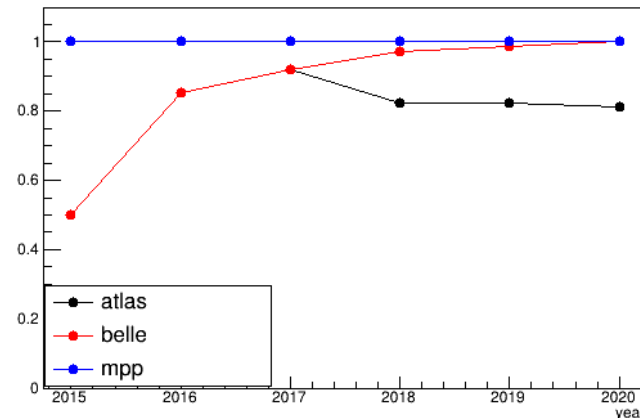
CPU server Henn and Zanderighi groups, 96 dual AMD EPYC 32 core

New storage (dCache)

CPU Shares

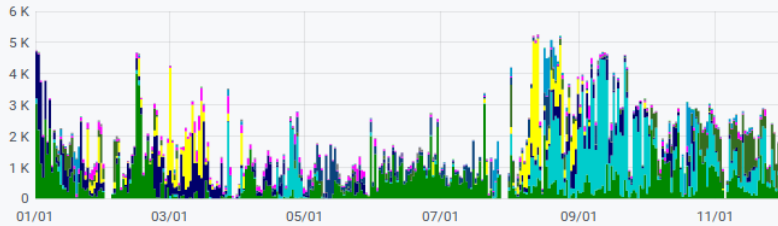


Storage Shares



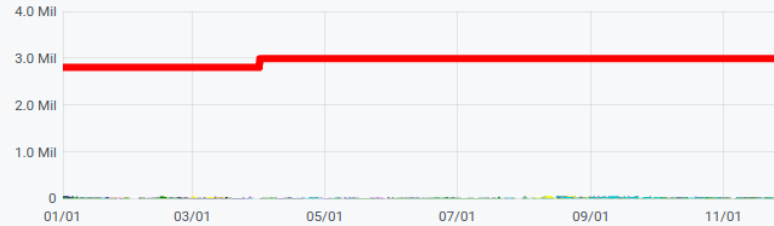
WLCG on MPP cluster @ MPCDF

Slots of Running jobs



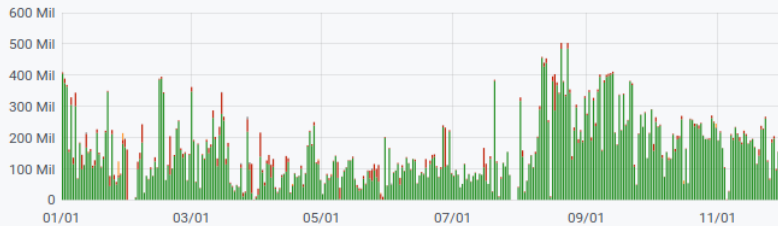
	min	max	avg	total
Group Production	0	3.996 K	544	182.357 K
MC Simulation Full	0	4.271 K	447	149.757 K
MC Reconstruction	0	2.609 K	258	86.596 K

Slots of Running jobs (HS06)



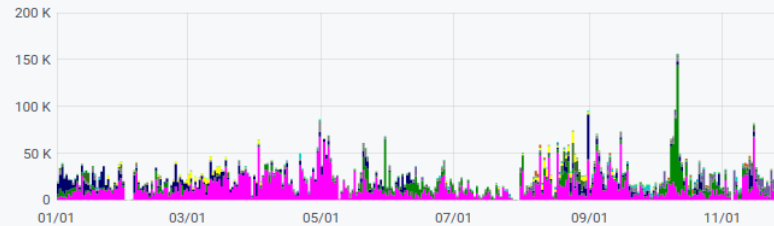
	min	max	avg	total
Pledges	2.799 Mil	2.985 Mil	2.935 Mil	3.926512 Bil
Group Production	0	45 K	6 K	2.052 Mil
MC Simulation Full	0	48 K	5 K	1.688 Mil

WallClock Consumption of Successful and Failed Jobs - Time Stacked Bar Graph



	min	max	avg	total
finished	0	485 Mil	152 Mil	50.768 Bil
failed	0	161 Mil	10 Mil	3.291 Bil
cancelled	0	44 Mil	363 K	121 Mil

Files processed



	min	max	avg	total
User Analysis	0	67.2 K	13.7 K	4.5820 Mil
Group Production	0	135.9 K	4.3 K	1.4473 Mil
MC Reconstruction	0	46.9 K	3.3 K	1.1137 Mil

Panda queues
MPPMU and
ANALY_MPPMU

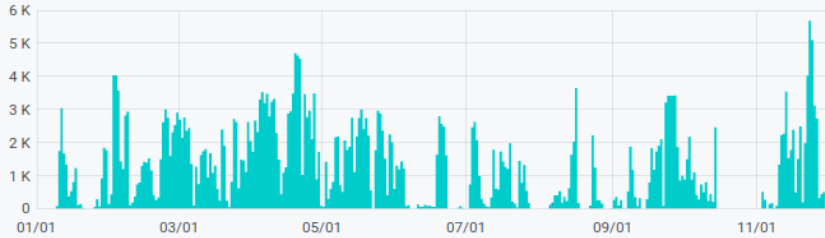
Pledges in HS06
wrong by factor
1000 ...

Almost there due to
Low WLCG share
In summer

WLCG on MPP cluster @ MPCDF

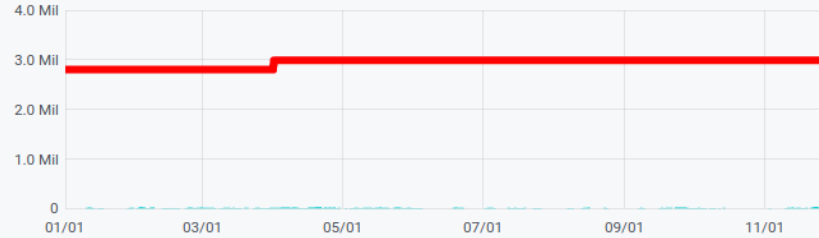
General

Slots of Running jobs



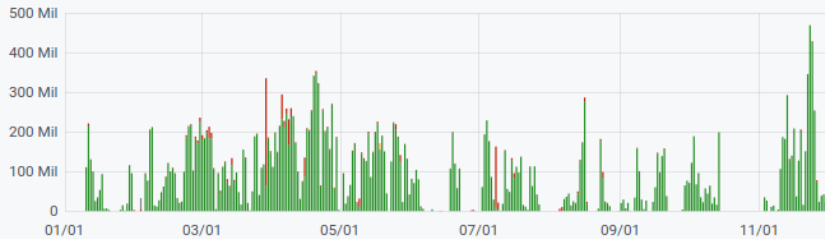
	min	max	avg	total
MC Simulation Full	0	5.685 K	1.162 K	389.144 K
MC Simulation Fast	0	103	1	184
Testing	0	1	0	4

Slots of Running jobs (HS06) MPPMU-DRACO_MCORE



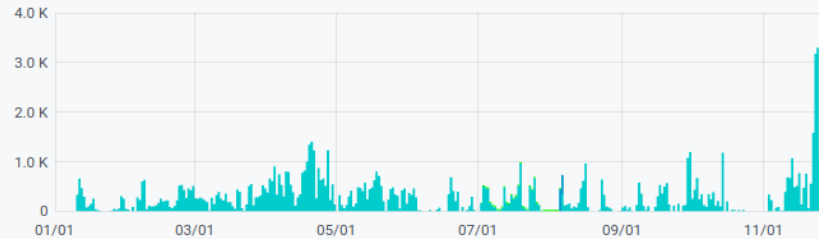
	min	max	avg	total
Pledges	2.799 Mil	2.985 Mil	2.935 Mil	3.926512 Bil
MC Simulation Full	0	31 K	6 K	2.076 Mil
MC Simulation Fast	0	548	3	983

WallClock Consumption of Successful and Failed Jobs - Time Stacked Bar Graph



	min	max	avg	total
finished	0	468 Mil	78 Mil	26.031 Bil
failed	0	270 Mil	3 Mil	1.116 Bil
closed	0	29 Mil	86 K	29 Mil

Files processed



	min	max	avg	total
MC Simulation Full	0	3.298 K	274	91.871 K
MC Simulation Fast	0	444	2	832
Testing	0	27	2	804

Examples from Theory group

Three-photon NNLO: MPCDF MPP cluster (20 x 5000 x 24h jobs) arxiv:2006.04133

Higgs, Drell-Yan $\text{MiNNLO}_{\text{PS}}$: MPP condor and MPCDF MPP cluster (ca. 50 x 2000 x 10h jobs, many tests+debug), arxiv:2010.04681

Z γ $\text{MiNNLO}_{\text{PS}}$: COBRA 100 x 1000 x 24h jobs, arxiv:2010.10478

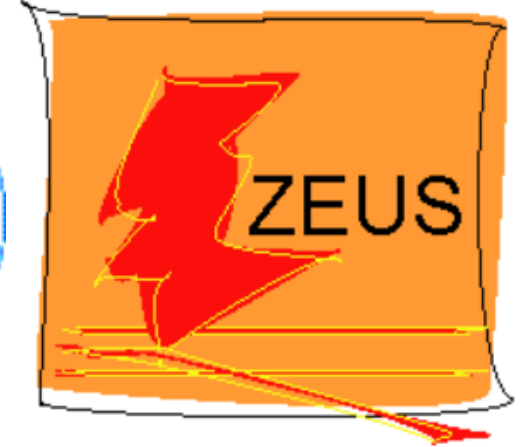
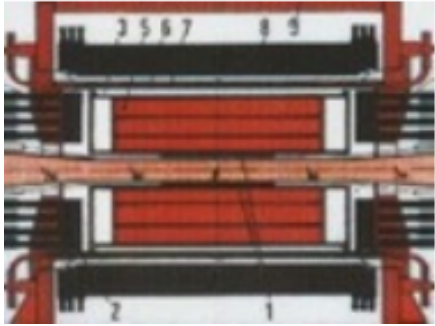
WW $\text{MiNNLO}_{\text{PS}}$: DRACO project ongoing, ca. 20 x 1000 x 12h jobs

ttbar $\text{MiNNLO}_{\text{PS}}$: MPP internet Condor Cluster (project ongoing ca. 30 x 500 x 10h jobs, Mazzitelli, Monni, Nason, Re, MW, Zanderighi)

Reduction to master integrals (FIRE6): zt 10 x 1-2 weeks jobs, 2-3 TB RAM, 60 cores, arXiv:2007.04851

Evaluation of millions of polylogs (GiNaC) to 15k digits: zt ca. 10 x 24h jobs, < 1 TB RAM, 60 cores, arXiv:2003.03120

Data preservation



e^+e^- annihilation: JADE (PETRA 1979-1985), OPAL (LEP 1989-2000)

ep scattering: H1, ZEUS (HERA 1990-2007)

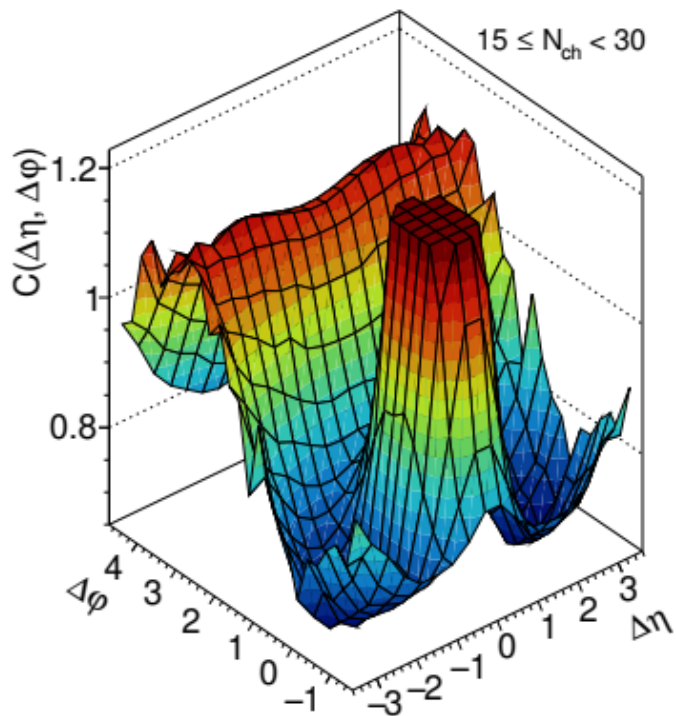
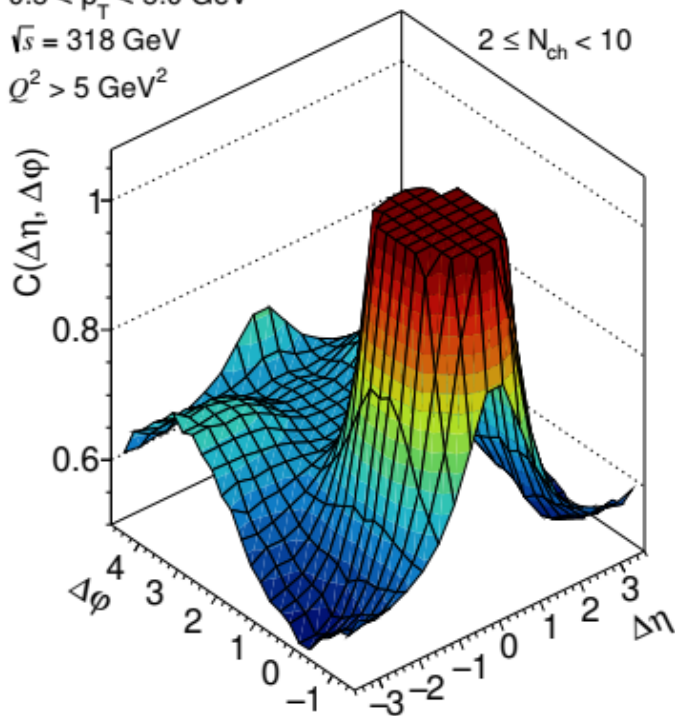
I. Abt^H, S. Bethke^{J,O}, D. Britzger^H, A. Caldwell^Z, V. Chekelian^H,
G. Grindhammer^H, J. Hessler^H, C. Kiesling^H, SK^{J,O},
A. Verbytskyi^{O,Z} (H. Abramowicz^Z, A. Levy^Z, H v.d.Schmitt^{J,O})

ZEUS two-particle correlations

Ridge effect in ep? $C(\Delta\eta, \Delta\phi) = N^{\text{pair}}_{\text{same}}(\Delta\eta, \Delta\phi) / N^{\text{pair}}_{\text{mixed}}(\Delta\eta, \Delta\phi)$

ZEUS

$0.5 < p_T < 5.0 \text{ GeV}$
 $\sqrt{s} = 318 \text{ GeV}$
 $Q^2 > 5 \text{ GeV}^2$



Near-side peak ($\Delta\phi \approx 0$)
Particles in same jet

Away-side peak ($\Delta\phi \approx \pi$)
Particles in other jet

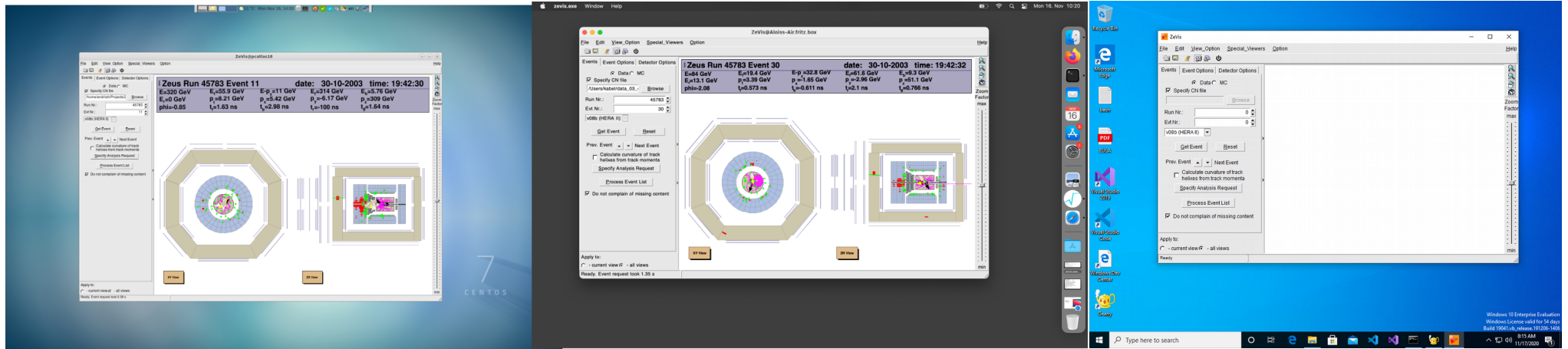
No evidence for extra
“hydrodynamic” two-
particle correlations in
ep (present in pp, pPb)

[JHEP 2004 (2020) 070]

DP machine room

- ZEUS software

- CNINFO (ZEUS DB): CentOS7/8, MacOS, W10
- FORMOZA (MC / HepMC3): CentOS8
- ZeVis (event display): C++ / Root6, cmake build, CentOS7/8, MacOS



JADE Software

- Fortran IV or 77 plus “extensions”
 - Late 70ies to 80ies
 - Now with gcc gfortran, or commercial Fortran compilers, cmake, Github CI, on CentOS7/8, MacOS
- Components
 - Ancient MCs (Jetset, Herwig 5.x, ...)
 - Detector simulation (w/o Geant!)
 - Reconstruction, event display, data reduction
 - Some more recent analysis codes

JADE, OPAL at CERN opendata

The screenshot shows the CERN Open Data Portal search results for 'JADE'. The browser window title is 'CERN Open Data Portal - Mozilla Firefox'. The search bar contains 'JADE' and the URL is '0.0.0:5000/search?page=1&size=20&experiment=JADE'. The search results are displayed in a list format, sorted by 'Best match' in ascending order. The results include:

- JADE Software: How to install
- JADE Software for reconstruction
- About JADE
- JADE Virtual Machines: How to install
- JADE author list
- Getting Started with JADE Open Data
- JADE computing notes

The left sidebar shows filters for type, experiment, year, file type, and keywords. The 'JADE' experiment is selected, showing 6 results. The 'Documentation' type is selected, showing 6 results. The 'Getting Started' type is selected, showing 1 result.

Data, software,
Documentation

JADE not fully
“public”, need
agreement of
collaboration

Could host in
similar way at
MPP (data al-
ready on own-
cloud)

SW deployment: COPR

COPR: "Community projects"

copr.fedorainfracloud.org/coprs/averbyts/fastjet

Supports CentOS7/8, Fedora, Suse, ...



Basis for stand-alone Singularity/Docker containers w/o LCG over /cvmfs, or for Github/Gitlab CI

Monitor averbyts/fastjet - Mozilla Firefox

Monitor averbyts/fastjet x +

https://copr.fedorainfracloud.org/copr/150% Search

fedora.copr™ log in | sign up

Search projects by name, os or arch

Home » averbyts » fastjet » Monitor » Simple

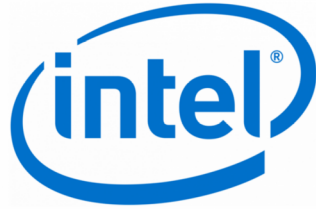
averbyts / fastjet
Project ID: 27759

Overview Packages Builds Modules Monitor

Build Monitor

	Epel 7	Epel 8	Fedora 33	Fedora rawhide
Package	x86_64	x86_64	x86_64	x86_64
applgrid	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
ariadne	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
astyle	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
BAT	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
blackhat	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
cascade	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
cernlib	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
CGAL	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
chaplín	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
clhep	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
collier	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
cuba	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
Delphes	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
DIRE	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
EvtGen	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
f2c	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
fastjet	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
fastnlo	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
fjcontrib	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded
form	✔ succeeded	✔ succeeded	✔ succeeded	✔ succeeded

Software preservation



vs.



Virtual machines, Linux Containers

Archive software and environment for later unmodified running. Risks: compatible hardware or VM / linux container environment disappear

Need proper build on new platforms: cmake, CI tests, packaging (COPR), automatic deployment

Just an example, Apple Si very competitive, don't expect every server to be on Apple Si anytime soon, but rapid transition to ARM possible

Summary

- Scientific computing at MPP and MPCDF
 - Central for many theory and exptl results
 - Increasing demand for CPU and storage
 - MPP: unify Linux, /remote/ceph (> 2PB) service, condor
 - MPCDF: new CPU server, > 5 PB storage
- Data preservation
 - Many physics results (only one example shown)
 - Rely on software (and documentation) preservation