Summary of the Performance session at the WLCG Workshop

Andrea Valassi (IT-DI-LCG)

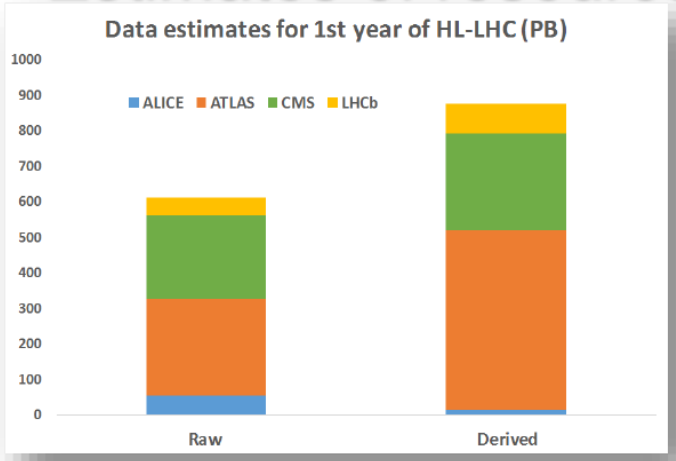
WLCG GDB – 9th November 2016

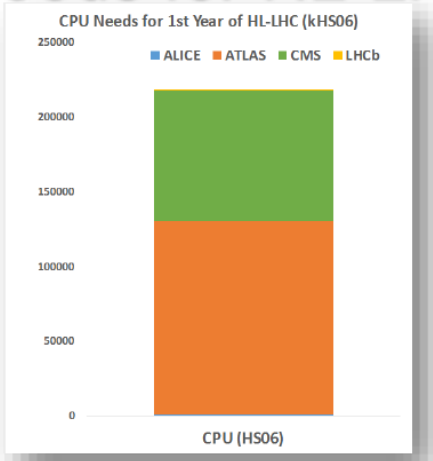
Performance talks at the WLCG Workshop

- Dedicated session (1.5 hours) on the 2nd Workshop day (Sunday morning)
- Talks from each experiment and from activities in CERN IT (agenda)

Introduction to the Performance session	*Andrea Valassi et al.*	
Hotel, San Francisco Marriott Marquis	09:00 - 09:05	
Workflow efficiency studies in CERN IT	*Gerhard Ferdinand Rzehorz*	
Hotel, San Francisco Marriott Marquis	09:05 - 09:15	
Performance studies in CERN IT	*Andrea Valassi*	
Hotel, San Francisco Marriott Marquis	09:15 - 09:20	
LHCb workflow efficiency studies	*Philippe Charpentier*	
Hotel, San Francisco Marriott Marquis	09:20 - 09:35	
ALICE workflow efficiency studies	*Maarten Litmaath*	
Hotel, San Francisco Marriott Marquis	09:35 - 09:40	
ALICE software performance studies	*Maarten Litmaath*	
Hotel, San Francisco Marriott Marquis	09:40 - 09:50	
ATLAS workflow and software performance studies	*Andrej Filipcic*	
Hotel, San Francisco Marriott Marquis	09:50 - 10:10	
CMS workflow and software performance studies	*David Lange*	
Hotel, San Francisco Marriott Marquis	10:10 - 10:30	

Estimates of resource needs for HL-LHC





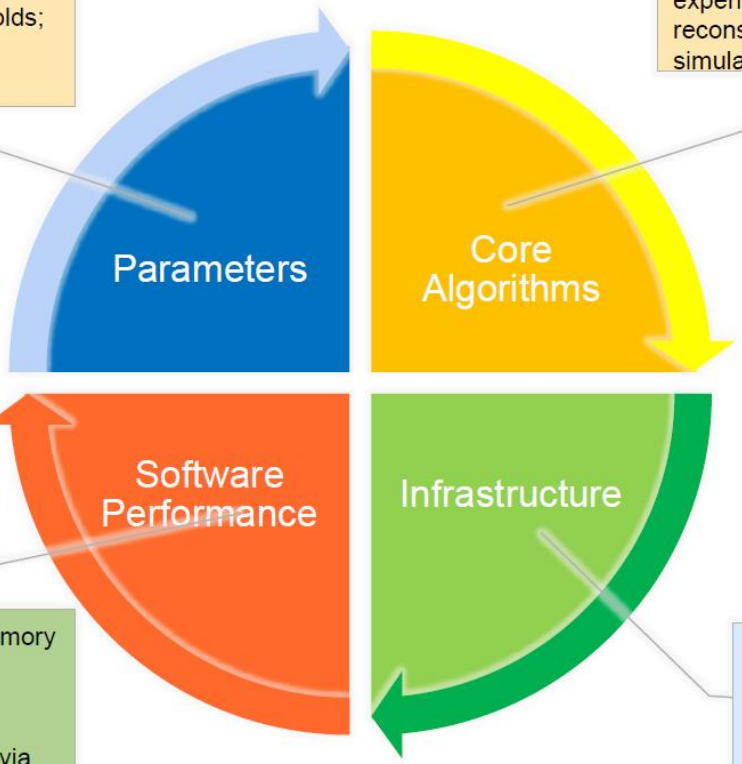
Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:
- x60 from 2016

Technology at ~20%/year will bring x6-10 in 10-11 years

- Simple model based on today's computing models, but with expected HL-LHC operating parameters (pile-up, trigger rates, etc.)
- At least x10 above what is realistic to expect from technology with reasonably constant cost

HL-LHC computing cost parameters



CERN | WLCG

8 October 2016 Ian Bird 11

A. Wiebalck – Reducing the performance penalty of VMs over bare metal

The “Kilo-1” configuration

- NUMA + Pinning
  - 1-to-1 vs. 1-to-N no difference
- 2MB huge pages
  - 1GB slightly better
- EPT on
  - EPT off still better in HS06

VM sizes (cores)	Before	After
4x 8	7.8%	3.3% (batch WN)
2x 16	16%	4.6% (batch WN)
1x 24	20%	5.0% (batch WN)
1x 32	20.4%	3-6% (bare SLC6 … batch WN)

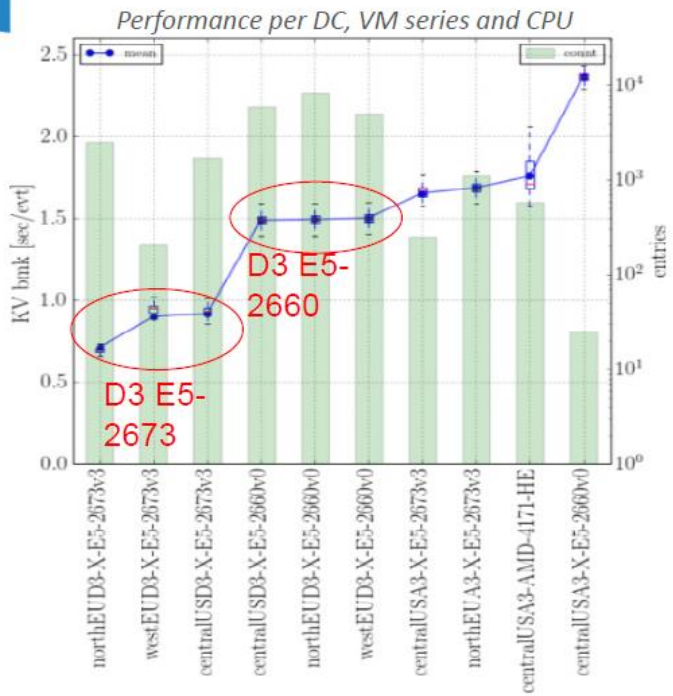
ATLAS T0 host with batch VM running the new config: throughput for recon jobs 20% higher!

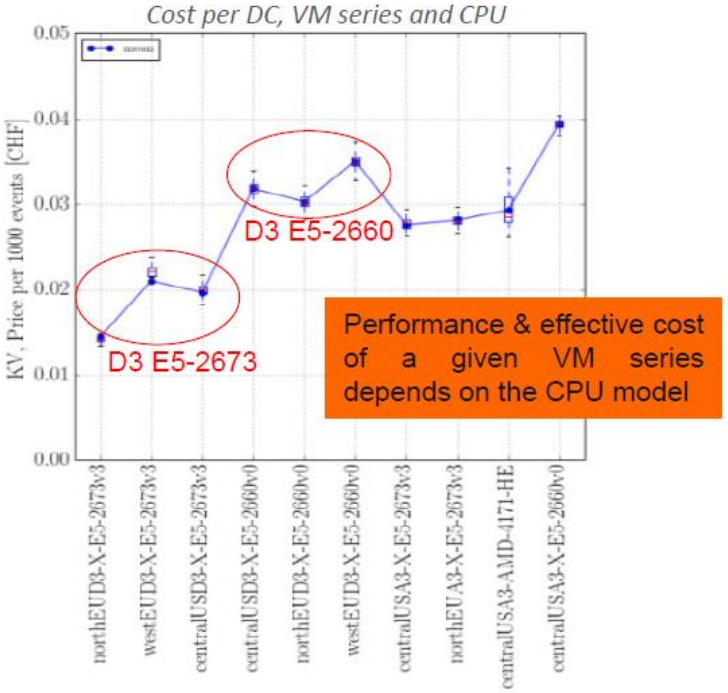
OpenStack Kilo will fully support our desired configuration!

G. Rzehorz – Workflow efficiency studies in CERN IT

C. Cordeiro – benchmarking commercial clouds integrated in WLCG

Deploy & *monitor & account & benchmark*



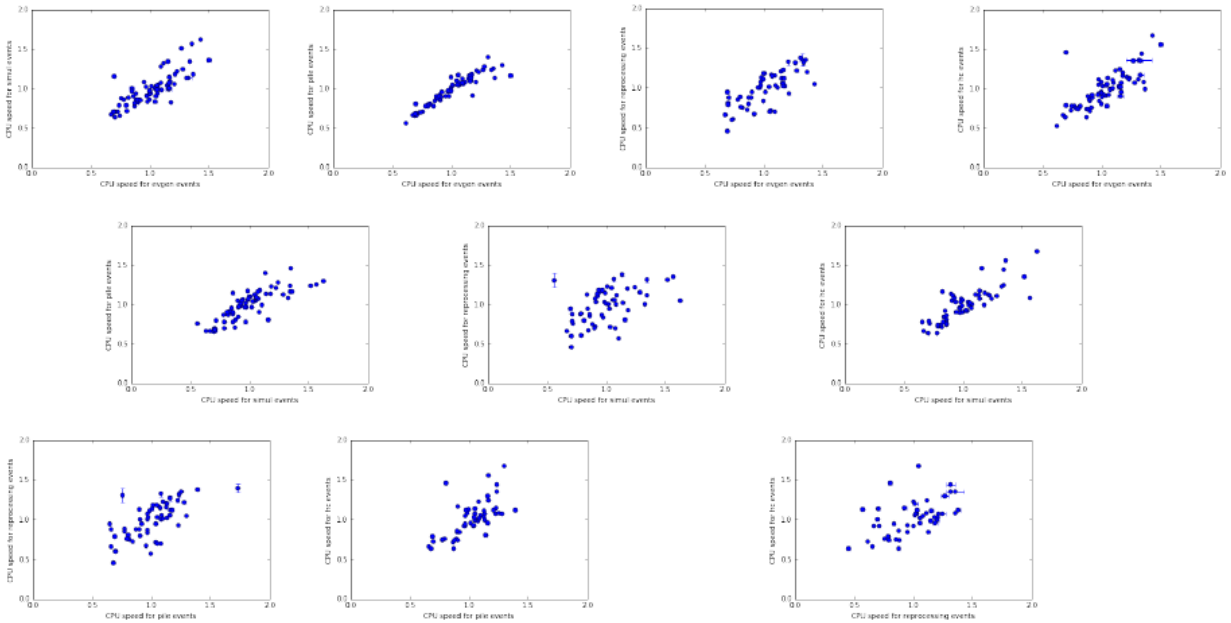


9

G. Rzehorz – Workflow efficiency studies in CERN IT

A. Sciaba – Fitting CPU speeds for different categories of ATLAS production jobs

Speed correlations



When considering only the three main categories (evgen, simul, pile), the speed factors agree by ~4% on average

Workflow and infrastructure activities

- Analysis of ATLAS jobs on the HLT farms
  - I/O patterns and constraints, workload behavior over time, OS and VM tuning
- Analysis of ATLAS Panda logs (being expanded also to CMS logs)
  - Differences between sites and between job types
  - Focus on understanding workflow efficiency
  - Fit CPU "speed" using data analytics and compare to benchmarks
- Set-up of a small cluster of 16 physical machines for performance studies
  - Batch "standard" nodes, complementary to Openlab/Techlab "exotic" nodes
- Evaluation of computing on storage servers (Andrei Kiryanov)
  - Most WLCG storage servers have low CPU utilization
- Prediction of available network capacity (Hendrik Borras, Marian Babik)

A. Filipcic (ATLAS) – Workflow performance meetings

ATLAS Workflow Performance Understanding chat

- Started informal ATLAS Computing Performance Understanding chat in May
- Meetings every 2-4 weeks with attendance from ADC, SW and CERN IT
- Discuss and understand ATLAS SW performance
- Understand bottlenecks and possible improvements and optimization in SW, ADC and grid infrastructure
- Make recommendations for workflow adaptations, grid workflows and hardware (how and where to run)

People: Andrej Filipcic, Markus Schulz, Andrea Sciaba, Andrea Valassi, David Smith, Johannes Elmsheuser, Sami Kama, Oliver Keeble, Alessandro Di Girolamo, Jaroslav Guenther, Tomas Kouba, Antonio Limosani, Nathalie Rauschmayr, Gerhard Ferdinand Rzehorz ...

- Mailing list: atlas-computing-workflow-performance@cern.ch

2

Where do we lose efficiency

Pilots:

- **Retirement** / draining of multicore pilots. If we split 8-core pilots into 8 single-core payloads, we must wait for the last single-core payload to end.
- **Negotiation**: Time between jobs and time-to-first-match.
- **Failed validation**: Pilot starts in unusable environment.

Payloads:

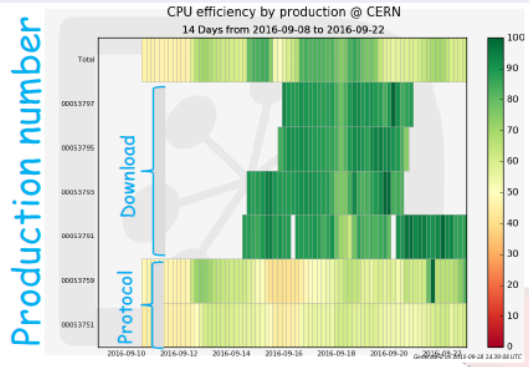
- **Job startup / shutdown**: Stage-in and -out. Exacerbated by short-running jobs
- **CPU efficiency**: Lots of IO per job (typically DIGI with pileup).
- **Algorithm performance**: Tasks that take longer than they should

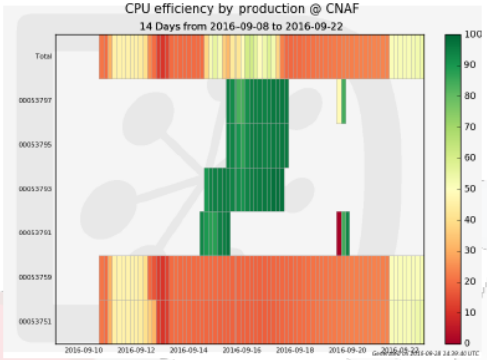
3

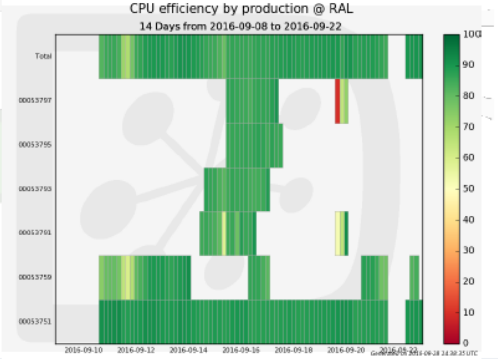
P. Charpentier (LHCb) – CPU efficiency depending on data access mode

Job CPU efficiency depending on access mode

WORKFLOW OPTIMISATION







- Job CPU efficiency includes all overheads
  - Startup, download if needed, output upload
- Bottom two productions: protocol access
- Next three productions: download access
- No doubt that downloading is preferable!

Improving analysis efficiency

- Users are advised to run big analyses as parts of analysis *trains*
  - Input data are read once and made available to all *wagons* of a train, each doing its own analysis
  - Train jobs have higher priority than individual analysis jobs
- The fraction of trains vs. individual analysis has strongly increased and then stabilized in the last years
  - 2015 Jan-Dec avg: 9k train jobs, 3.9k individual jobs, 61k total
  - 2016 Jan-*Sep* avg: 8k train jobs, 3.9k individual jobs, 73k total
    - The small decrease is due to the delayed reconstruction of the 2015 heavy-ion run, now mostly done
    - Since the start of autumn: 10k train jobs on average!
- Users are also advised to run analysis over compact AOD files instead of the 10x bigger, sparsely useful ESD files
  - Try to reduce unnecessary reading of large amounts of unused data
  - AOD analysis generally has higher priority than ESD analysis

CPU Efficiency improvements on T0

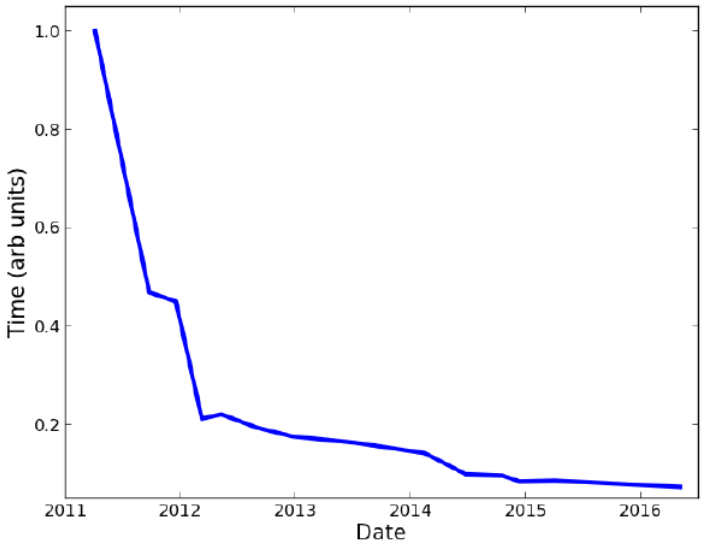
	Tier0 2015	1st repro	2nd repro	Tier0 2016
Release	20.1.8.3	20.7.3.5	20.7.3.8	20.7.6.4
Cores	1	8	8	1
Max PSS [GB]	3.2	14.7	13.1	3.6
Avg PSS [GB]	1.5	6.5	4.5	2.9
Max Swap [GB]	3.3	0.008	0.008	1.8
Avg Swap [GB]	0.4	0.0	0.0	0.007
Avg CPU eff	≈ 0.6	0.55	0.46	≈0.95
Avg CPUtime/evt [s] (RAWtoESD)	13.3	-	-	7.8
Avg Walltime/evt [s] (RAWtoESD)	17.2	-	-	8.1
Avg Walltime/evt [s] (all)	-	3.7	5.7	-
Avg Walltime/evt [s] (corr.)		12.0	18.1	

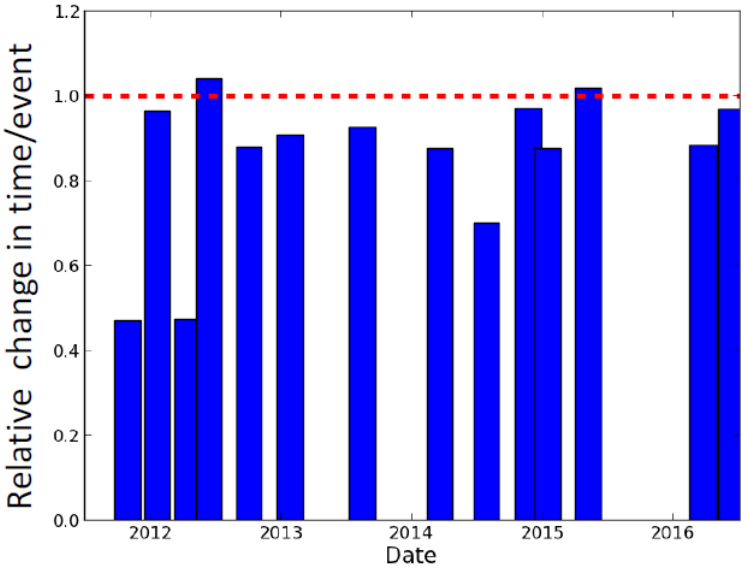
- Tuning the data processing on Tier0 resulted in ~95% efficiency
  - Significant improvement since 2015 at ~60%

- data15 13TeV.00282712.physics Main.daq.RAW (2015 run)
- data16 13TeV.000300415.physics Main.daq.RAW (2016 run)

10

Reconstruction performance against high pileup data





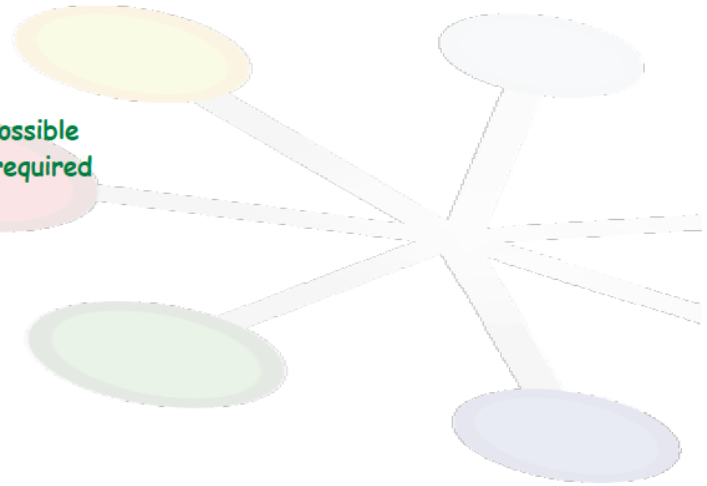
Incremental progress continues through Run 2 while maintaining the physics

8

WORKFLOW OPTIMISATION

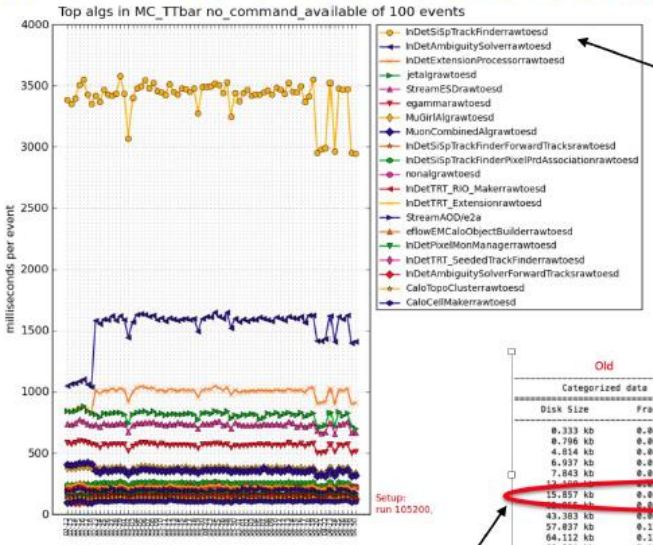
What I will not talk about…

- **Software optimisation**
  - There is a major effort that went into this already
  - Reconstruction speed improved by a factor 2 (20 HS06.s/event to 10 HS06.s/event)
    - During LS1
  - Simulation: speedup
    - Use parametrized showers whenever possible
    - Avoid simulating RICH photons if not required
    - …
  - Major effort for the LHCb upgrade
    - See CHEP presentation
    - Complete re-engineering of
      - Software framework (Gaudi)
      - Event model
      - Computing Model
- **In this talk**
  - Concentrate on how to most efficiently run the existing software



A. Filipcic (ATLAS) – Software profiling analysis

CPU and data-size consumption by components



Top consumer of CPU is Tracking/Inner Detector, but everything else matters too!

Old

Categorized data		
Disk Size	Fraction	Category Name
0.333 kb	0.001	*Unknown*
0.796 kb	0.002	MetaData
4.014 kb	0.009	Muon
6.937 kb	0.013	EvtId
7.843 kb	0.015	MET
15.857 kb	0.031	tau
43.383 kb	0.083	BTag
57.037 kb	0.110	PFO
64.112 kb	0.123	Jet
68.881 kb	0.133	CaloTopo
95.344 kb	0.184	Truth
108.000 kb	0.208	InDet
519.570 kb	1.000	Total

New

Categorized data		
Disk Size	Fraction	Category Name
0.333 kb	0.001	*Unknown*
0.803 kb	0.002	MetaData
4.014 kb	0.009	Muon
6.937 kb	0.013	EvtId
7.842 kb	0.015	MET
13.198 kb	0.025	CaloTopo
22.102 kb	0.042	tau
43.383 kb	0.083	BTag
57.037 kb	0.108	PFO
64.112 kb	0.122	Jet
68.881 kb	0.131	CaloTopo
95.344 kb	0.181	Truth
108.000 kb	0.206	InDet
525.820 kb	1.000	Total

Size matters too — even small size increases can become significant on the grid

20

M. Litmaath for S. Wenzel (ALICE) – Software performance improvements

Overall Results

- Improvements measured on running a Pb-Pb simulation (typical event) - Geant3
- **~20% gain** in total runtime achieved (from ~2359s to ~1966s)

	Original	Tuned compiler flags	Code optimizations in AliRoot/ROOT
RunSimulation	1462s	1367s	1182s
RunSDigitization	683s	692s	585s
Total simulation + all digitization + other parts	2359s	2274s	1966s

Sandro Wenzel WLCG workshop, San Francisco, 8-9/10/2016 6

Software-related activities

- Analysis of memory usage/dynamics and development of related tools
  - FOM-tools (Find Obsolete Memory), x32-ABI re-evaluation
- Evaluation of tracing tools (SystemTap), studies on CPU hardware counters
  - Aggregate knowledge about other existing tools (Coverity, VTune...)
- Studies on Feedback-Directed Optimization
  - Job profiles helping compilers to auto optimize code (AutoFDO)
- Investigating differences between Intel microarchitectures
  - IPC (instructions per cycle) on Haswell vs Ivy Bridge for some simulation jobs
  - Analyzing effects at the level of instruction caches and branch prediction units
- Analysis of a few specific software components (e.g. Sherpa MC generator)

Also: performance-related talks at CHEP

- Personal selection of some of the talks/posters I attended and/or found interesting
  - A. Forti, Memory Handling in the ATLAS Submission System (link)
  - P. Calafiura, Tracking Machine Learning Challenge (link)
  - A. Perez, CMS Readiness for Multi-Core Workload Scheduling (link)
  - C. Haen, Monitoring Performance of LHCb Computing Infrastructure (link)
  - S. Kama, Identifying Memory Allocation Patterns in HEP Software (link)
  - S. Y. Jun, Computing Performance of GeantV Physics Models (link)
  - S. Campana, The ATLAS Computing Challenge for HL-LHC (link)
  - C. Bozzi, The LHCb Software and Computing Upgrade for Run3 (link)
  - G. Stewart, How to Review 4 Million Lines of ATLAS Code (link)
  - O. Keeble, Combined Analysis of CERN Computing Infrastructure Metrics (link)
  - T. Limosani, Monitoring of Computing Resource Use in ATLAS (link)
  - D. Abdurachmanov, Investigation of Future Computing Platforms for HEP (link)
  - T. Childers, Challenges in Scaling NLO Generators to Leadership Computers (link)
  - S. Chapeland, A Programming Framework for Data Streaming on XeonPhi in ALICE (link)
  - D. Riley, Kalman Filter Tracking in Parallel Architectures (link)
  - C. Gumpert, From ATLAS Software towards “A Common Tracking Software” (link)
  - S. Wenzel, Accelerating Navigation in the VecGeom geometry model (link)
  - P. Hobson, Parallel Monte Carlo Search for Hough Transform (link)
  - P. Hristov, Blurring Online and Offline (link)
  - P. Conde Muino, GPUs in the ATLAS HLT (link)
  - D. Rohr, GPU Accelerated Track Reconstruction in ALICE HLT (link)
  - F. Pantaleo, Accelerated Tracking Using GPUs for CMS HLT in Run3 (link)
  - D. Campora, LHCb Kalman Filter Cross-Architecture Studies (link)
  - S. Farrell, Multi-Threaded ATLAS Simulation on Intel Knights Landing (link)
  - P. Charpentier, Benchmarking Worker Nodes Using LHCb Simulation Productions (link)
- I will NOT cover the talks above in this presentation!

Also: performance-related talks at CHEP

- Personal selection of some of the talks/posters I attended and/or found
  - A. Forti, Memory Handling in the ATLAS Subsystems
  - R. Gelfi

Conclusions and take-away message:
A LOT of the ongoing work concerns PERFORMANCE!

Optimization of software
Optimization of workflow/infrastructure/dataaccess
Investigation of radically new platforms and approaches

Within each experiment
Within infrastructure teams
In collaboration across experiments and common teams

- ...NOT cover the talks above in this presentation!