Karlsruhe Institute of Technology

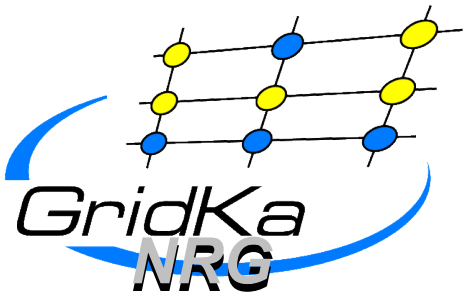
Batch system data locality via managed caches

Max Fischer
HEPiX spring meeting 2014

Steinbuch Centre for Computing / Institute of Experimental Nuclear Physics

Background – KIT LHC Computing

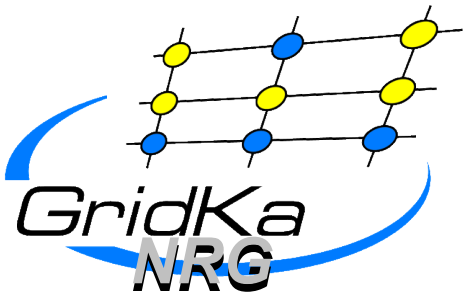
- Campus North: GridKa
  - Multi-VO Tier1
    - See Talks by M. Alef, E. Kühn
  - German users exclusive „National Resources at GridKa”
    - Analysis user access to Tier1 CPU share
    - Access to T1 dCache
    - Dedicated NRG dCache
    - Deliberate usage only by power users

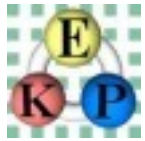


Steinbuch Centre for Computing
Institute of Experimental Nuclear Physics

Background – KIT LHC Computing

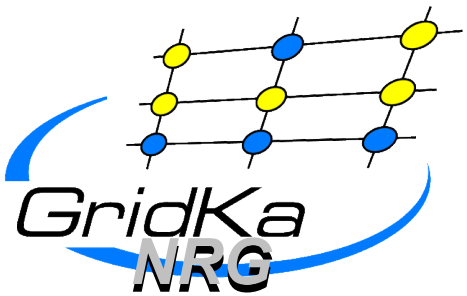
- Campus North: GridKa
  - Multi-VO Tier1
  - German user exclusive „National Resources at GridKa”
- Campus South: EKP (CMS) resources
  - Former T3_DE_Karlsruhe
    - Local SGE workgroup cluster (~200 cores)
    - Local storage (~200TB)
  - Experimental processing resources
    - HPC/Cloud resources via GlideinWMS
    - HTCondor Desktop VM cluster

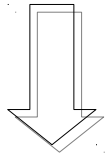


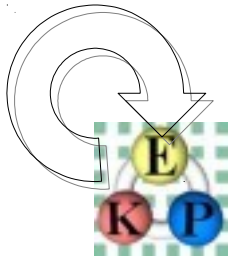


KIT CMS analysis user workflows

- Low-Frequency official data pre-processing
  - Grid based, semi-automated processes
  - Skimming official data sets from grid storage
  - Output stored at NRG dCache
- Manual transfer of data to EKP fileservers
  - Faster access from EKP network
  - POSIX :)
- High frequency analysis data processing
  - Local batch cluster, constantly evolving processes
  - Reading current private data set from fileservers
  - Output stored at file servers



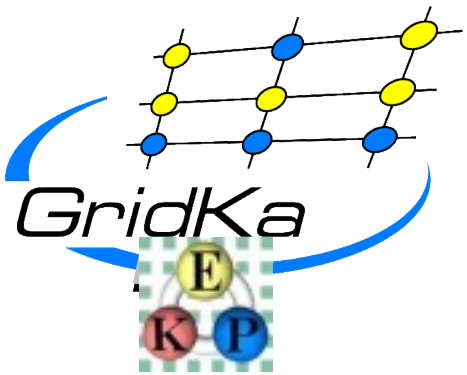




Steinbuch Centre for Computing
Institute of Experimental Nuclear Physics

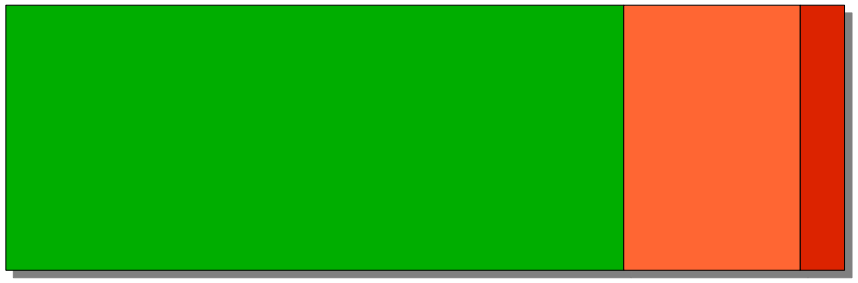
Joined Tier3 infrastructure

- Combining data generation and processing
  - Extending EKP cluster into NRG
  - Providing direct NRG dCache access via pNFS
  - Technically not a problem
- Deal Breaker: Data Rates
  - Aiming for data input intense workflows
  - High clock speed CPUs in EKP desktops
  - NRG Storage access performance limited



User analysis data profile

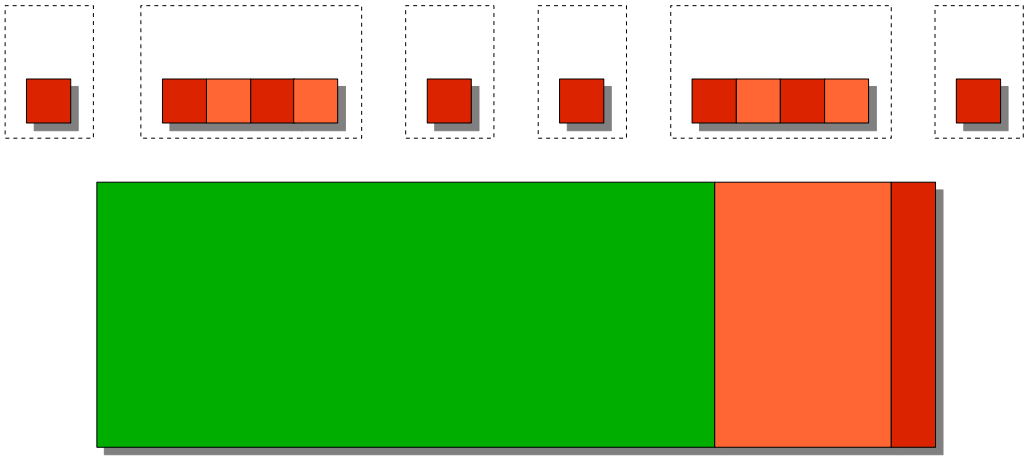
- Only fraction of user data is “active”
  - ~1 TB “daily usage” data
  - 2-8 TB overall physics data
  - ~50 MB source code
  - ~1 MB non-recoverable configuration
- Example: CMS Jet Energy Corrections power user
  - 0.5 TB data set processed on daily scale
  - ~2 TB data set processed on weekly scale
  - ~7 TB data set processed quarterly



Steinbuch Centre for Computing
Institute of Experimental Nuclear Physics

Managed data locality

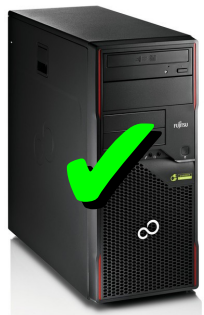
- All data in storage/file servers
  - Complete set of physics data and files
  - GridKa dCache & EKP Servers
  - What we have already
- Most regularly used files cached on workers
  - Full data locality for most common workflows
  - Caches coordinated for best coverage
  - Fallback to network access



Steinbuch Centre for Computing
Institute of Experimental Nuclear Physics

Raw HW Benchmarking (compressed)

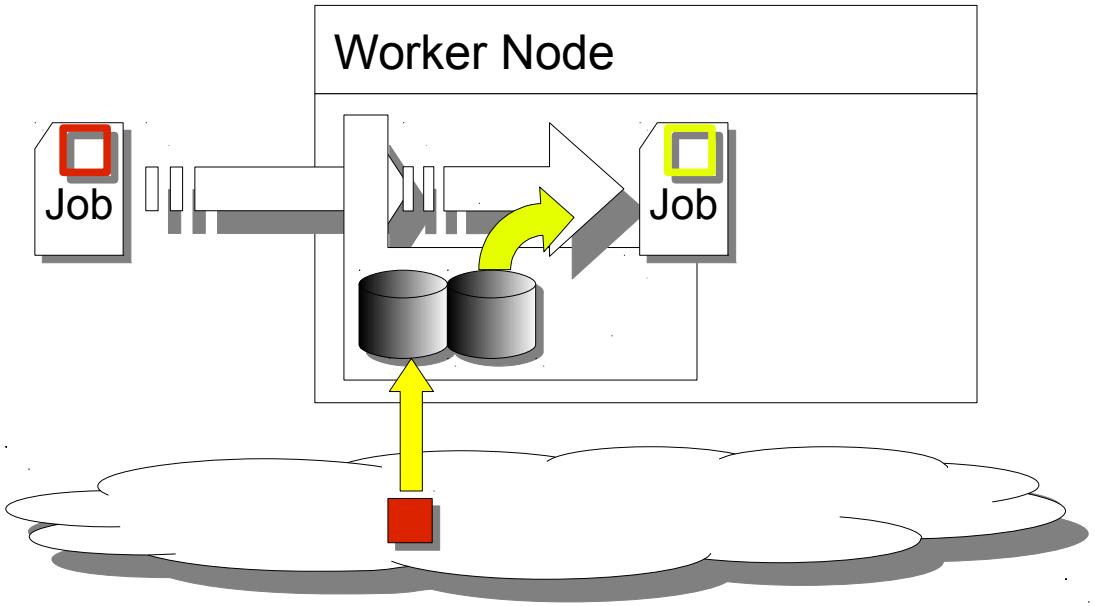
- Server WN with caches
  - Did not get one in time :(
- Desktop WN with caches
  - SSD Cache limits
    - Theoretical read speed >> Achievable analysis throughput
    - Virtually no loss from concurrent access
    - Capacity/$ still poor
  - HDD Cache limits
    - Acceptable read speed
    - Concurrent access read speed degradation



Steinbuch Centre for Computing
Institute of Experimental Nuclear Physics

Concept - WN

- Cache janitor process
  - Check cache consistency with storage
  - Fetch missing/outdated files from storage
- Expose cache for jobs and scheduling
  - Export cache content list as WN details
  - Point job to cache by modifying config via job wrapper



Concept - Pool

- Track file access patterns in batch system
  - User, frequency, task size, concurrency, …
  - Identify high-profile files and data
- Assign files to cache pool
  - Issue caching orders via batch system scheduling
  - Add/remove cached files from WN caches

Feedback & Questions

- Questions or Comments?

Steinbuch Centre for Computing
Institute of Experimental Nuclear Physics