Setup and commissioning of a high-throughput analysis cluster

Rene Caspart*, Florian von Cube, Max Fischer, Manuel Giffels, Christoph Heidecker, Andreas Heiss, Eileen Kühn, Andreas Petzold, Günter Quast

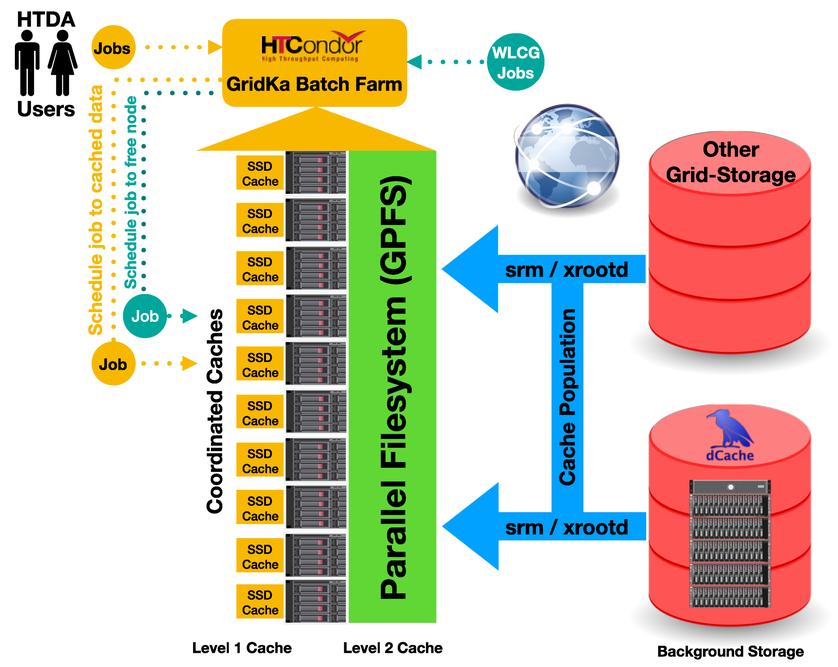
Introduction

Goals

- Set up an analysis focused Tier-3 cluster within a Tier-1 facility
- Profit from existing Tier-1 infrastructure
- Cluster optimized for usage with distributed hierarchical caching approaches
  ⇒ **T**hroughput **Op**timized **A**nalysis **S**ystem

Setup of the Cluster

- 11 hyperconvergent workernodes
- 1 TB NVMe and 96 TB HDD for caching per node
- 100 Gbit/s network connection and 200 Gbit/s uplink
- 1 GPU-node with 8 Nvidia V100

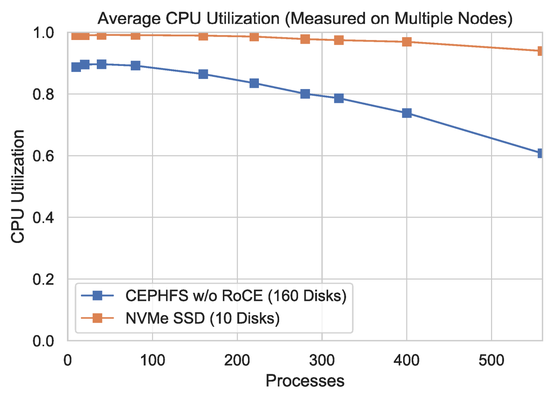


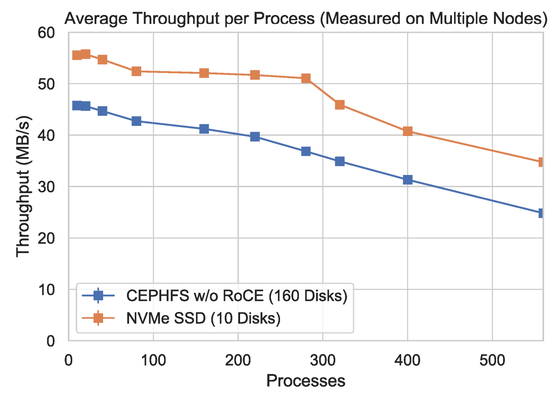
Benchmarks

Disk and throughput benchmark

- Testing analysis-like workflow
- Reading data from a ROOT-file
  ⇒ Typical speed limitation: ~50 MB/s and per core

- Benchmark with up to 560 cores on 10 nodes
- The benchmark is performed for two setups
  - Using NVMe SSDs
  - Using HDDs with CEPHFS as distributed filesystem





Almost ideal CPU utilization when reading from NVMe SSDs

Usage of the TOPAS Cluster

Institute resources and development cluster

- High-throughput extension of the institute batch cluster
  - User jobs are flocked to the TOPAS cluster
- Development cluster for caching approaches
  - Distributed caches
  - Hierarchical caches
    ⇒ See poster 510 by Max Fischer

Opportunistic usage

- Cluster often not fully utilized
- Backfilling with WLCG jobs
  - Jobs are running in preemptable slots
  - Using COBalD and TARDIS developed at KIT
    ⇒ See talks by Manuel Giffels and Max Fischer in Track 7 (Opportunistic resources) on Thursday

Rene Caspart, Steinbuch Centre for Computing: rene.caspart@kit.edu