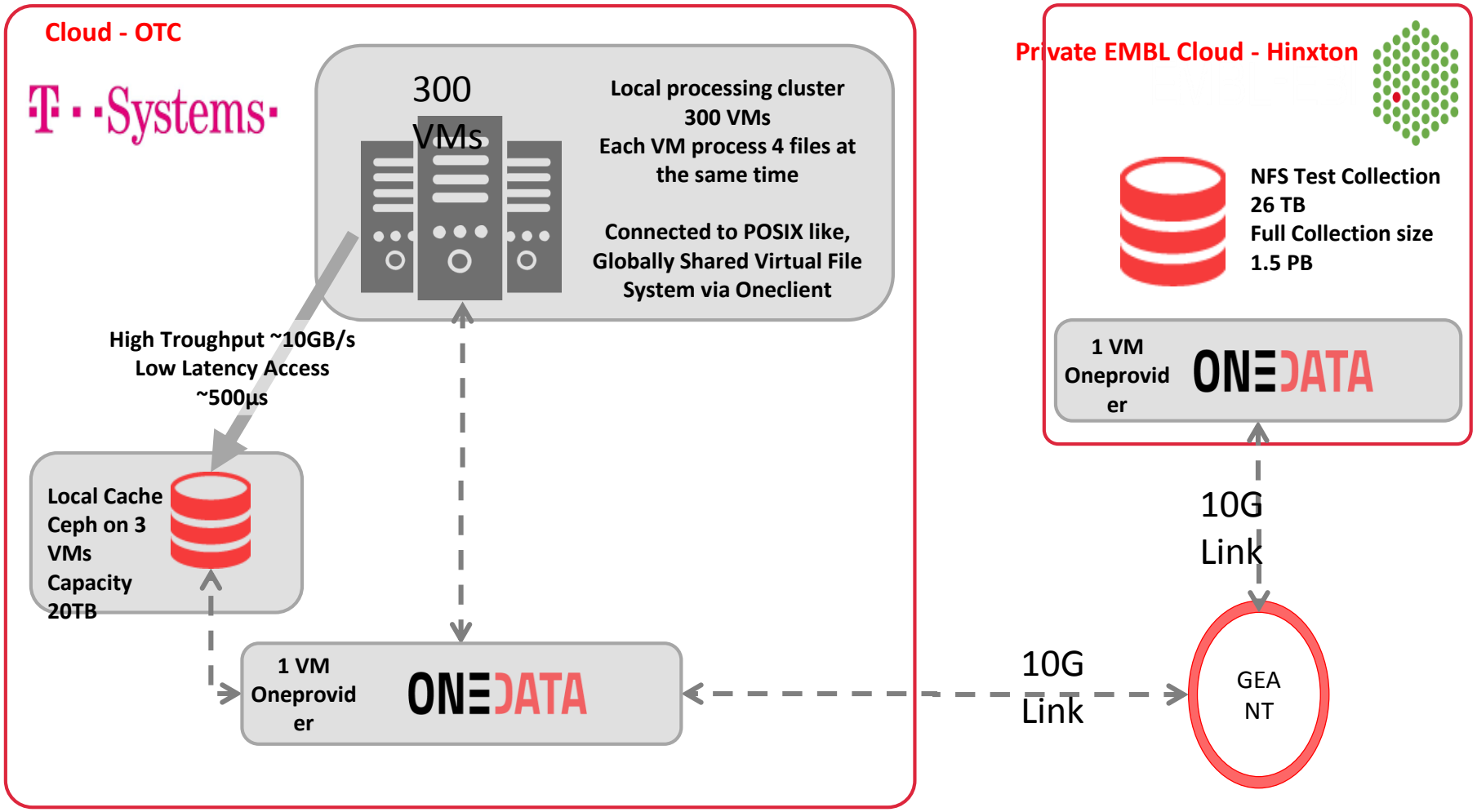
HNSciCloud

PanCancer update

Tony Wildish

EMBL-EBI

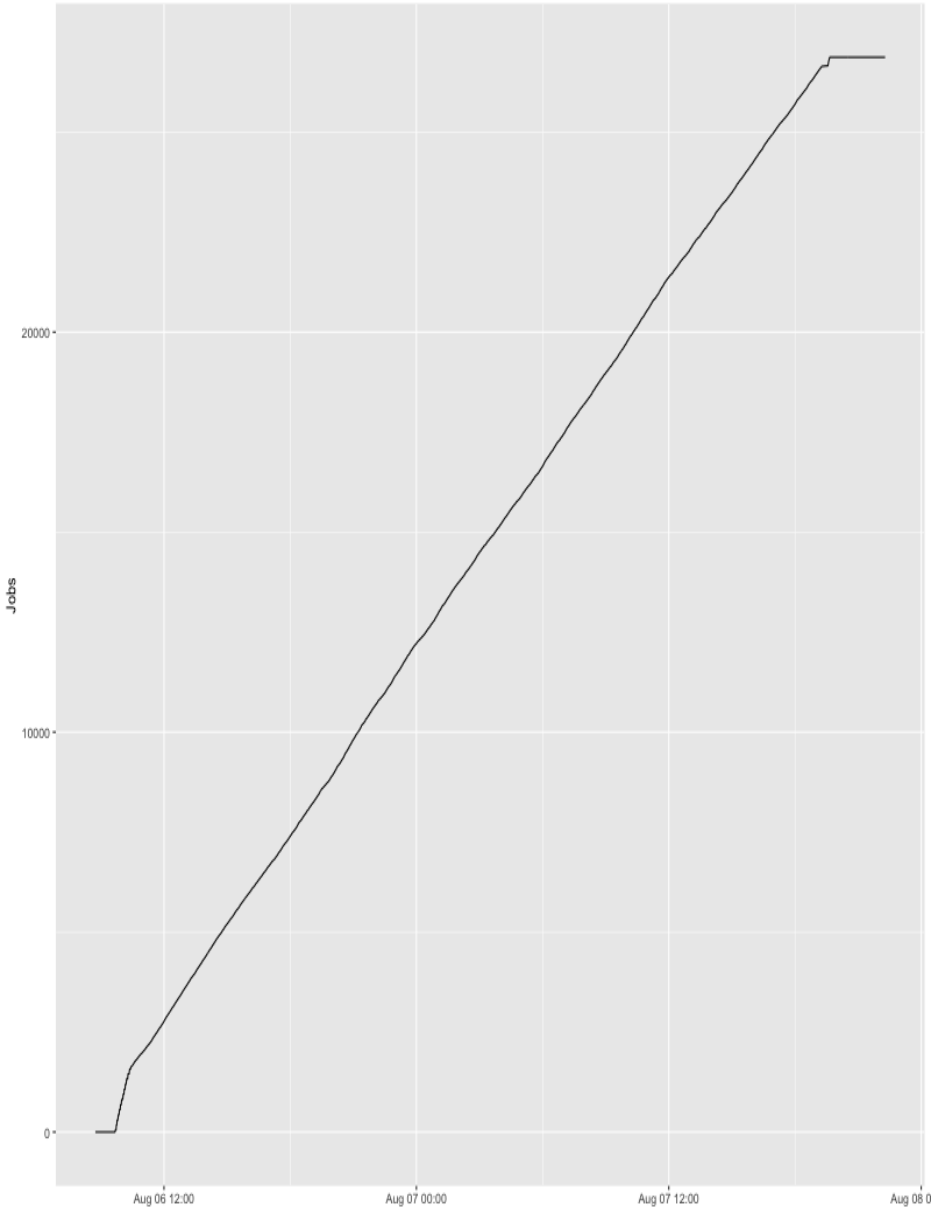
EMBL-EBI Application Deployment

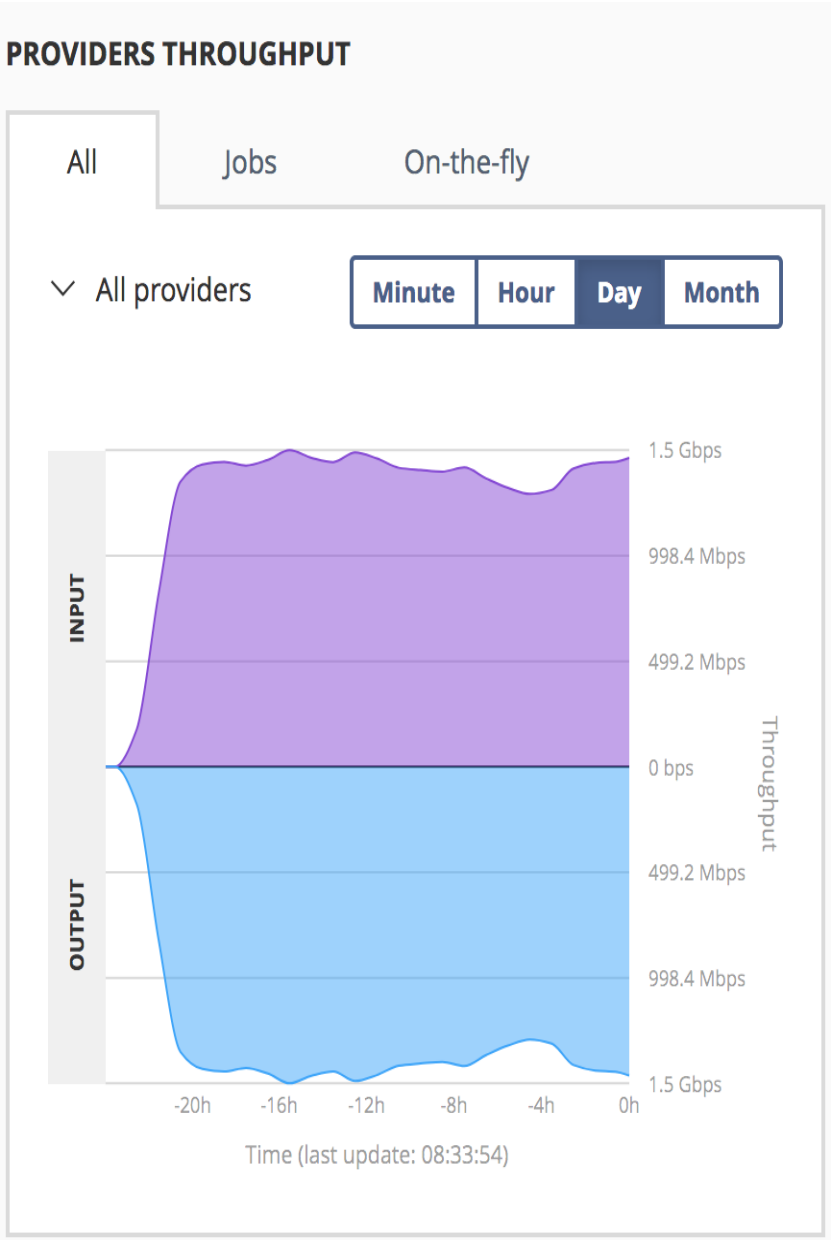


EMBL-EBI Application

- DNA Sequence alignment based on the Pan-Cancer project use-case
- Application requires POSIX access, no native object-store or WAN-aware version exists
- 1046 Input DNA sequences, one per file => ~15 – 50 GB per file
- Each input is compared in turn to each of 24 reference human chromosomes
  - 22 ‘normal’ chromosomes plus X & Y
  - ~25K batch jobs, each lasts ~minutes to several hours
  - 300 batch VMs, each running 4 tasks (single-core)
  - Entire run takes 1-2 days
- No local data on OTC before experiment starts
  - Each input file processed 24 times, reference genome used >1000 times => caching is very useful
- Pre-staging not needed, Onedata automatically delivers needed blocks and optimises access
  - Adaptive block-level pre-fetching, data delivered on the fly, automatic cache management

Job-throughput, data serving rates





Onedata test results

- Onedata is maturing rapidly
  - Multiple repeat runs show consistent, stable behavior over several weeks, good performance
  - Not all possible access-patterns tested, there are use-cases that haven't been explored
- Small overhead on runtime performance, extra cost from extra infrastructure
  - Tradeoff w.r.t. rapid cloud deployments with minimal change of application framework
- Cache tuning and optimization is an art, as always, but all the knobs and levers are there
  - How useful it will be depends on your use-case, i.e. how often and when you re-use your data
- Viable option for POSIX-access to remote filesystem
  - Avoids DDoS by managing client-access
  - Limits network exposure by reducing holes in firewalls