Karlsruher Institut für Technologie

Analyzing data flows of WLCG jobs at batch job level

Christopher Jung, KIT

STEINBUCH CENTRE FOR COMPUTING - SCC

Disclaimer

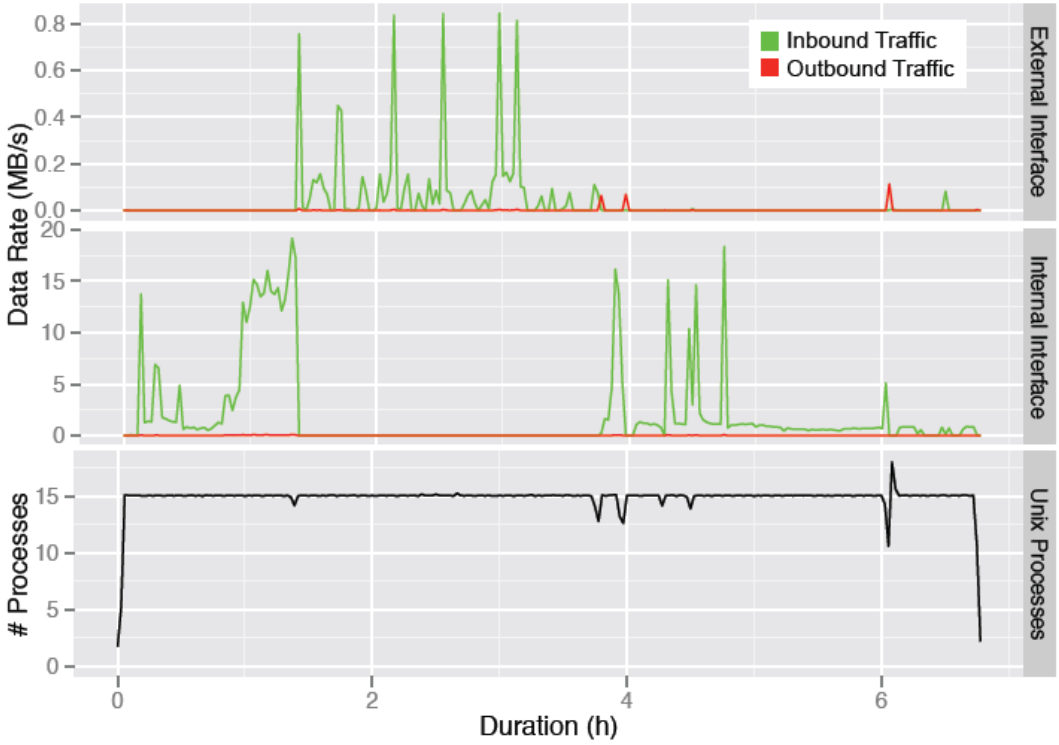
- The main author is Eileen Kühn, PhD student in Computer Science at KIT

- Currently, she is at CERN Summer School of Computing in Braga

Retrieved from http://commons.wikimedia.org/wiki/File:Braga_Montage.png, Creative Commons Attribution-Share Alike 2.0 Generic

Motivation

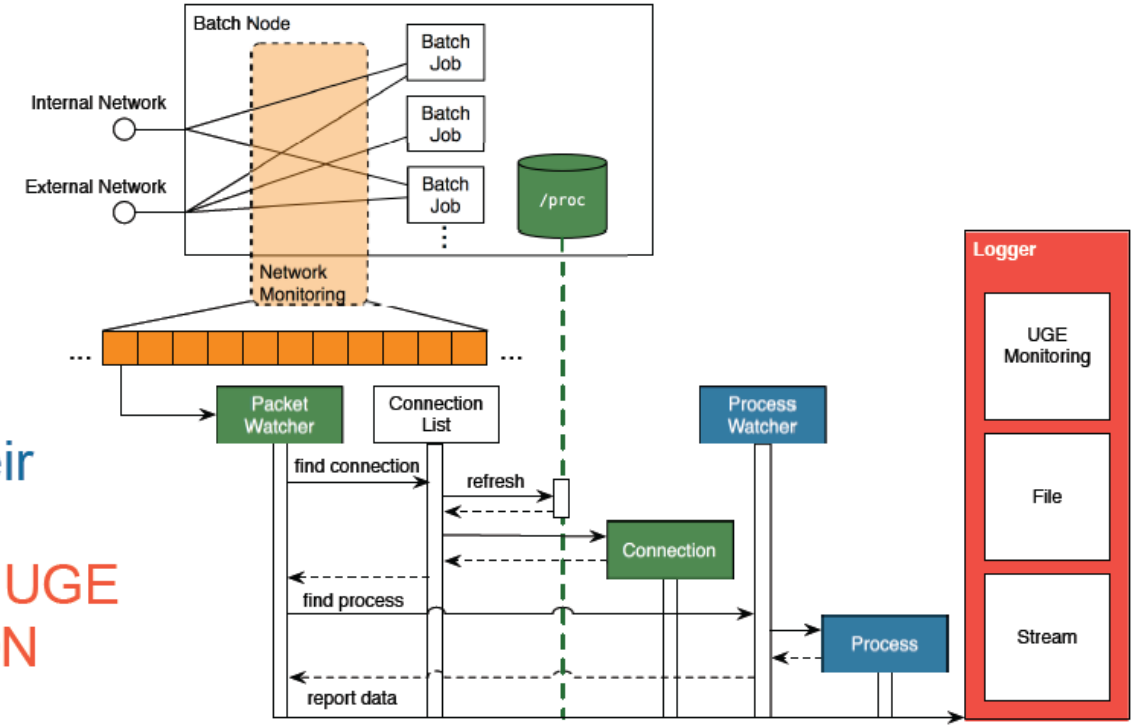
- Increase in complexity of data flows:
  - Federated data access
  - #cores/WN still growing
  - Pilot jobs
- Analysis of data streams on batch job level needed
  - Predict storage element access
  - Predict future resource requirements
  - Detect abnormal usage of resources



Realization

Job Monitoring

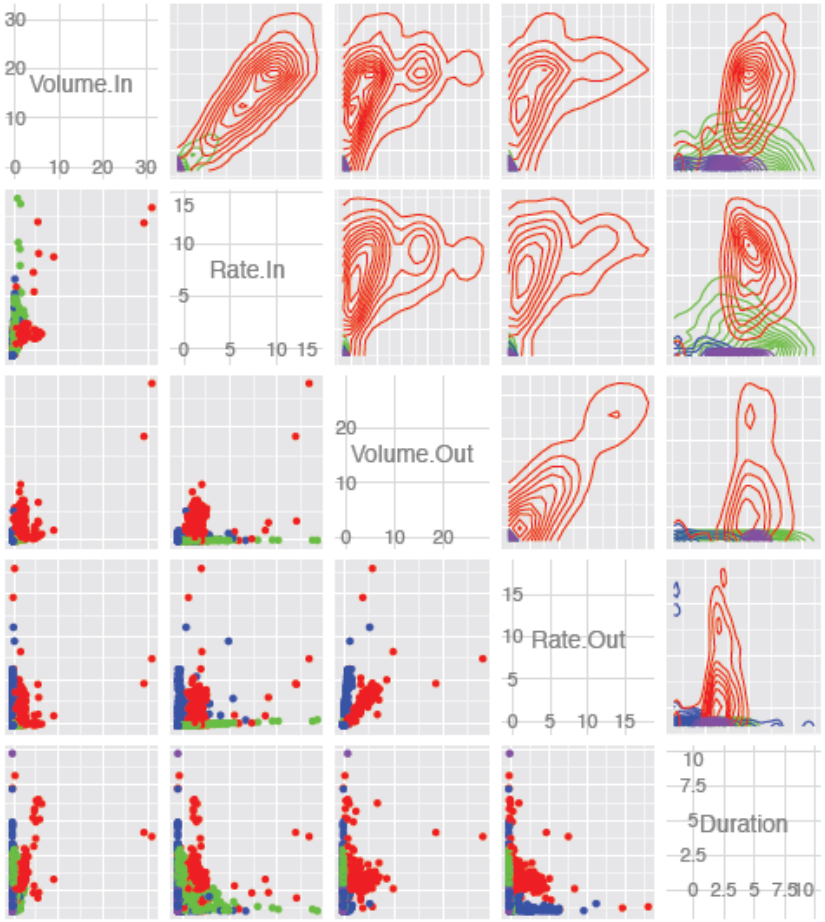
- Monitoring of UDP/ TCP packets
- Splitting into internal/ external network traffic
- Group processes by their associated batch jobs
- Output to file/stream as UGE command, CSV, or JSON



- Tools used for collecting information: /proc, libpcap, netfilter

Toy Analysis (I)

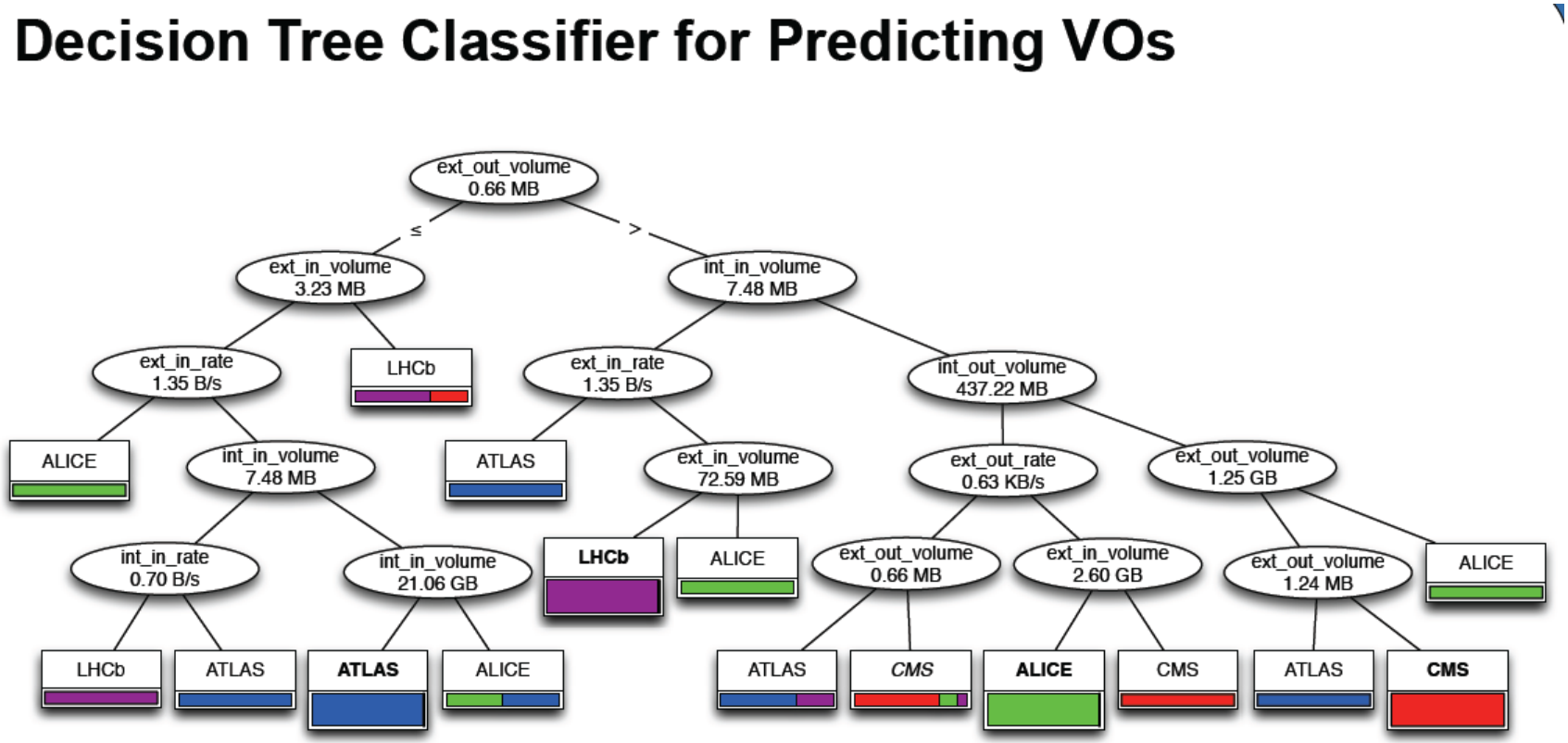
Question: How specific are data volumes, rates and duration per job for LHC VOs?



- ~3.800 jobs (duration>1h)
  - Values normalized
  - Data rates and volumes for internal network traffic

Toy Analysis (II)

Using internal and external volume and rates (in and out):



	ALICE	ATLAS	CMS	LHCb
Purity	98.0%	98.4%	93.1%	99.0%
Efficiency	95.4%	99.2%	95.7%	98.2%