Future of Batch Processing at CERN

HEPiX Fall 2013

Jérôme Belleman Daniel Pek
CERN IT

October 2013

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Outline

1 The Future, LSF and its Shortcomings

2 Alternative Batch Systems

3 Initial Results

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Section 1

The Future, LSF and its Shortcomings

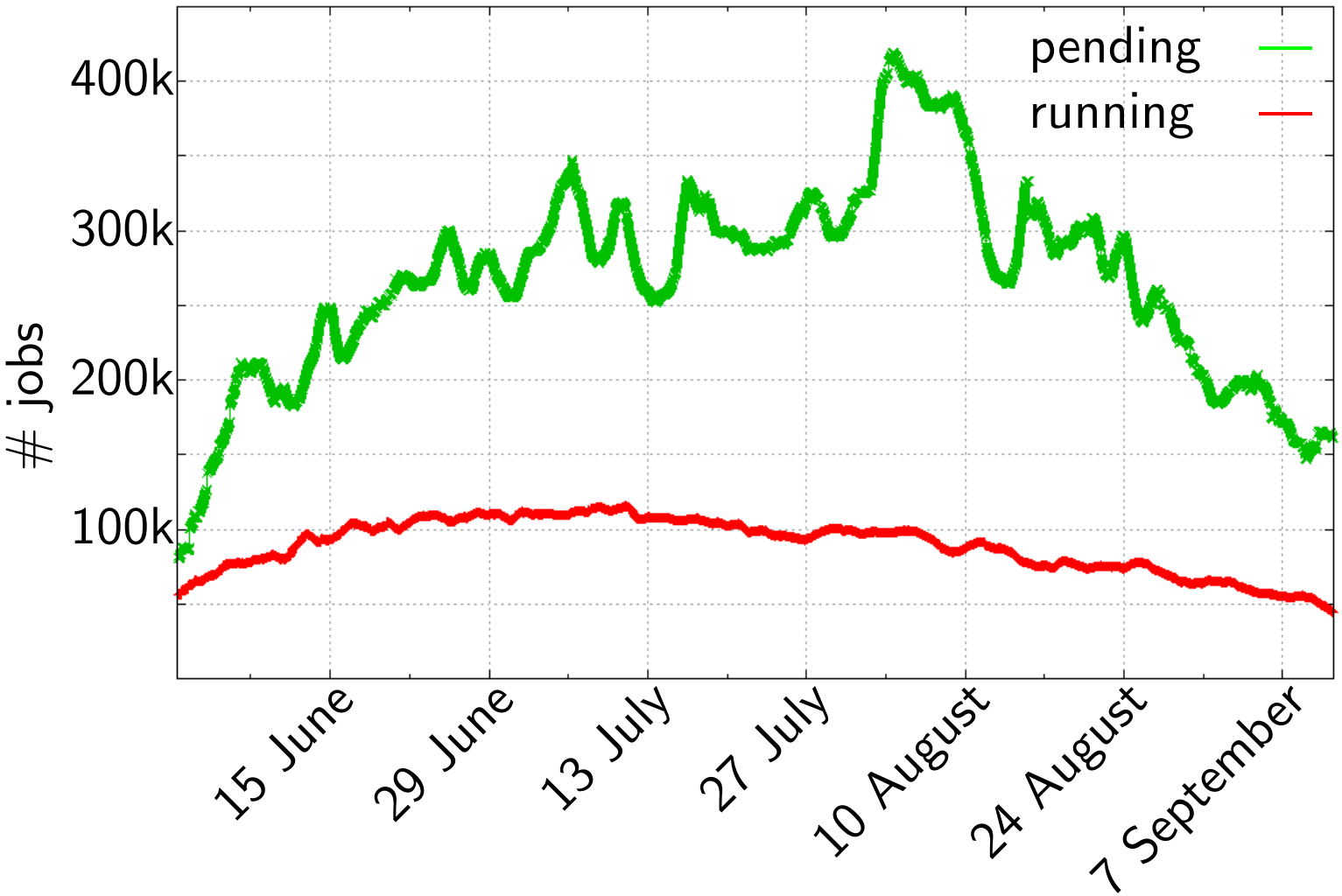
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Current Setup

- IBM LSF 7.0.6
- 4 000 nodes
  - SLC5 $\rightarrow$ SLC6
  - Physical $\rightarrow$ virtual ($\approx$ 1000 virtual so far)
- $>$ 65 000 cores
- 400 000 jobs/day
- $\pm$70 000 running jobs

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

A Large, Busy Cluster



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Goals and Concerns

Goals	Concerns with LSF
30 000 to 50 000 nodes	6 500 nodes max
Cluster dynamism	Adding/Removing nodes requires reconfiguration
10 to 100 Hz dispatch rate	Transient dispatch problems
100 Hz query scaling	Slow query/submission response times

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Addressing Operational Issues

- Hired an LSF consultant
- Suggested a few minor enhancements:
  - Kernel parameters
  - Memory limits
  - CPU binding
  - Logging and accounting file handling
- No miraculous improvements
- Taught us a lot

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Section 2

Alternative Batch Systems

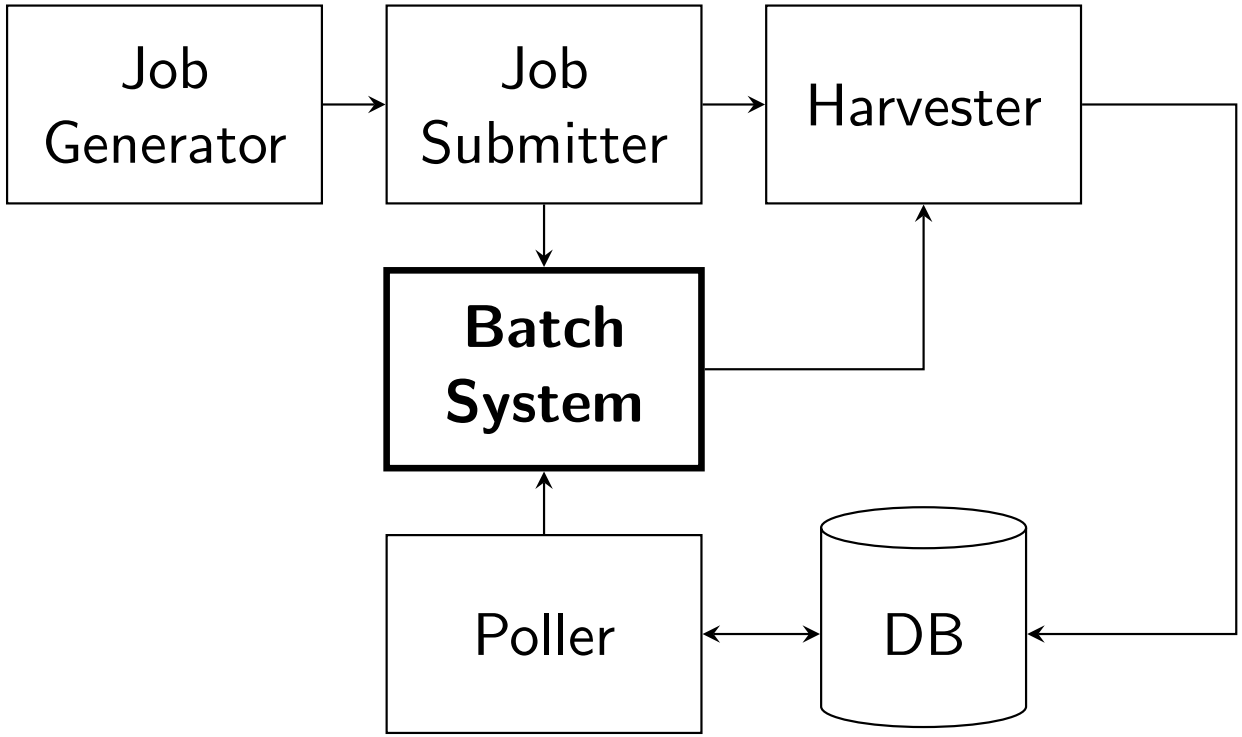
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Current Candidates

- SLURM 2.5.7
- HTCondor 8.1.0
- Son of Grid Engine 8.1.3
- LSF 8/9

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

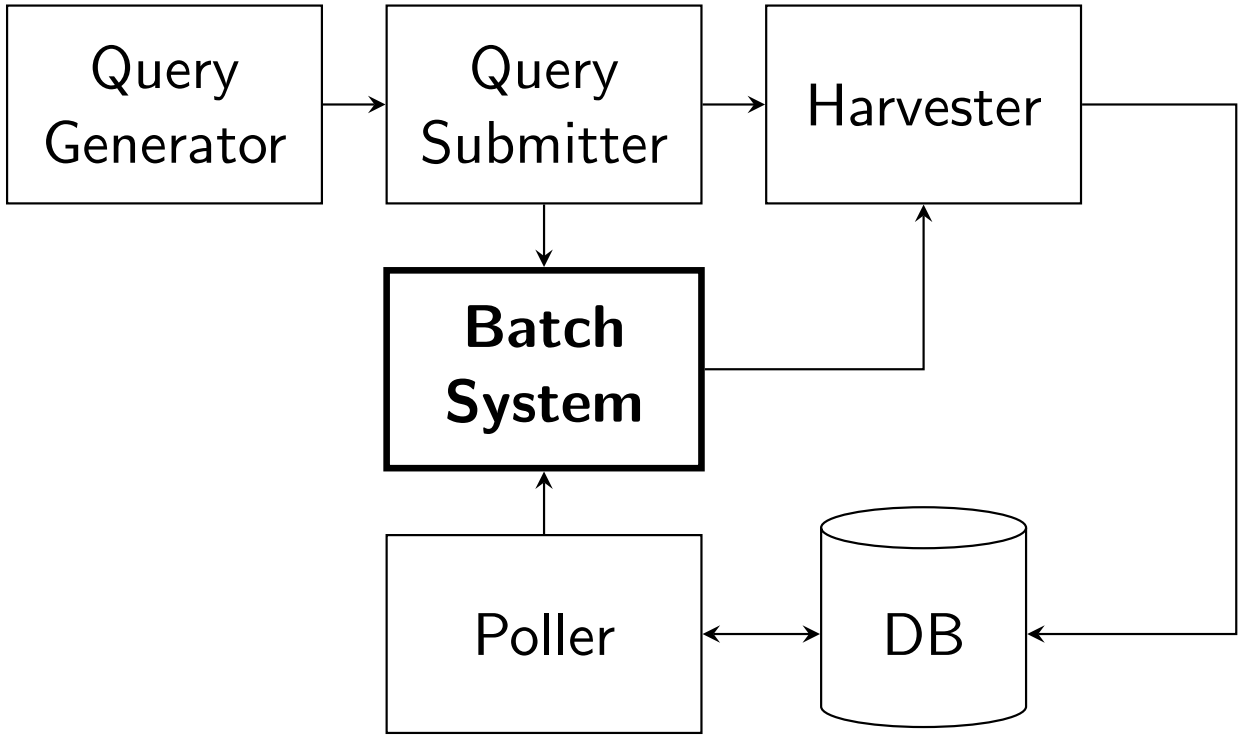
Batch Testing Framework




```
[
    { "cmd":       "tests.d/psleep.py 120",
      "count":     1000,
      "instance":  "batch",
      "user":      "atlas",
      "queue":     "debug" }
]
```

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Batch Testing Framework




```
[
    { "cmd":       "tests.d/bjobs.py",
      "count":     1,
      "instance":  "batch" }
]
```

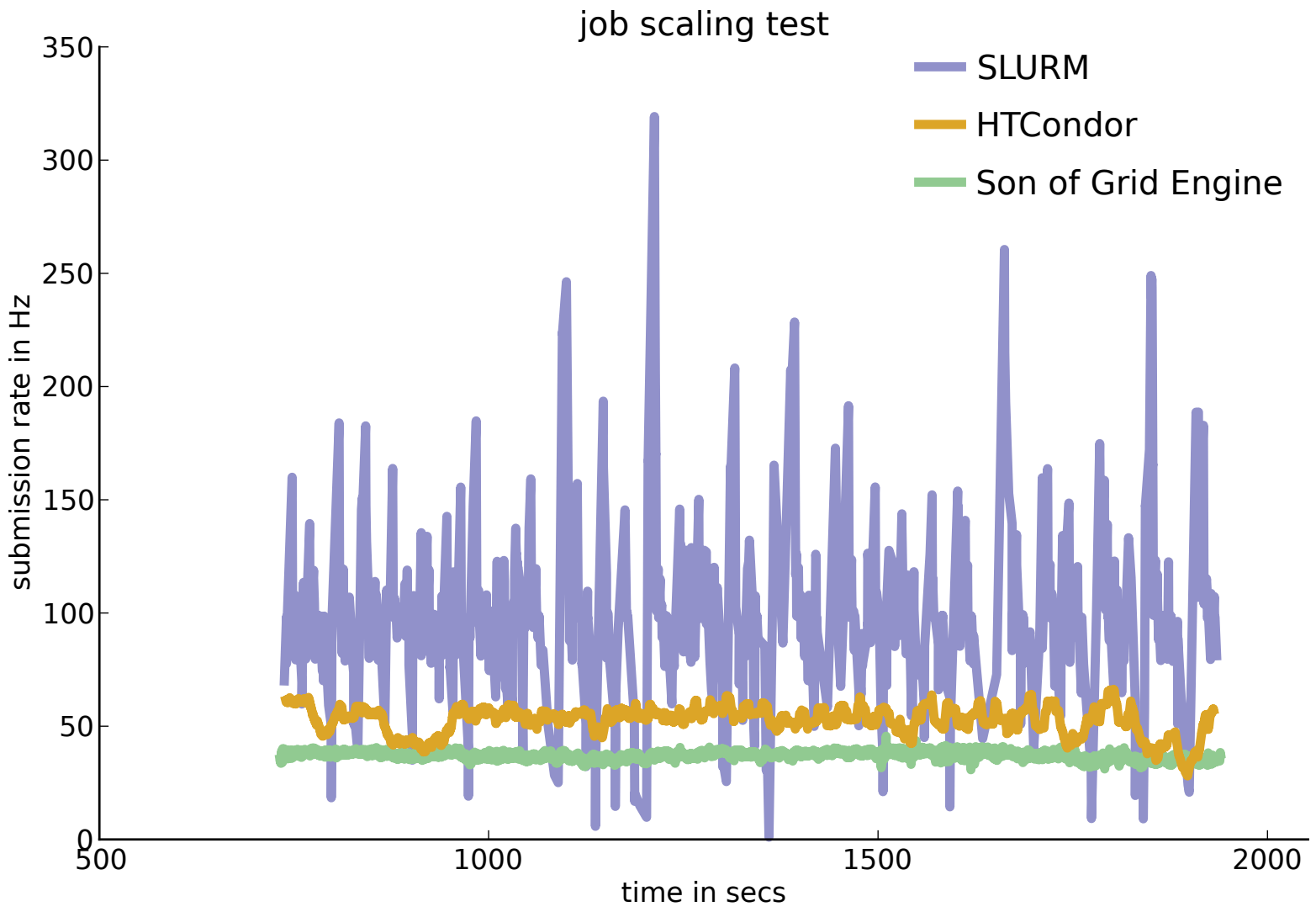
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Section 3

Initial Results

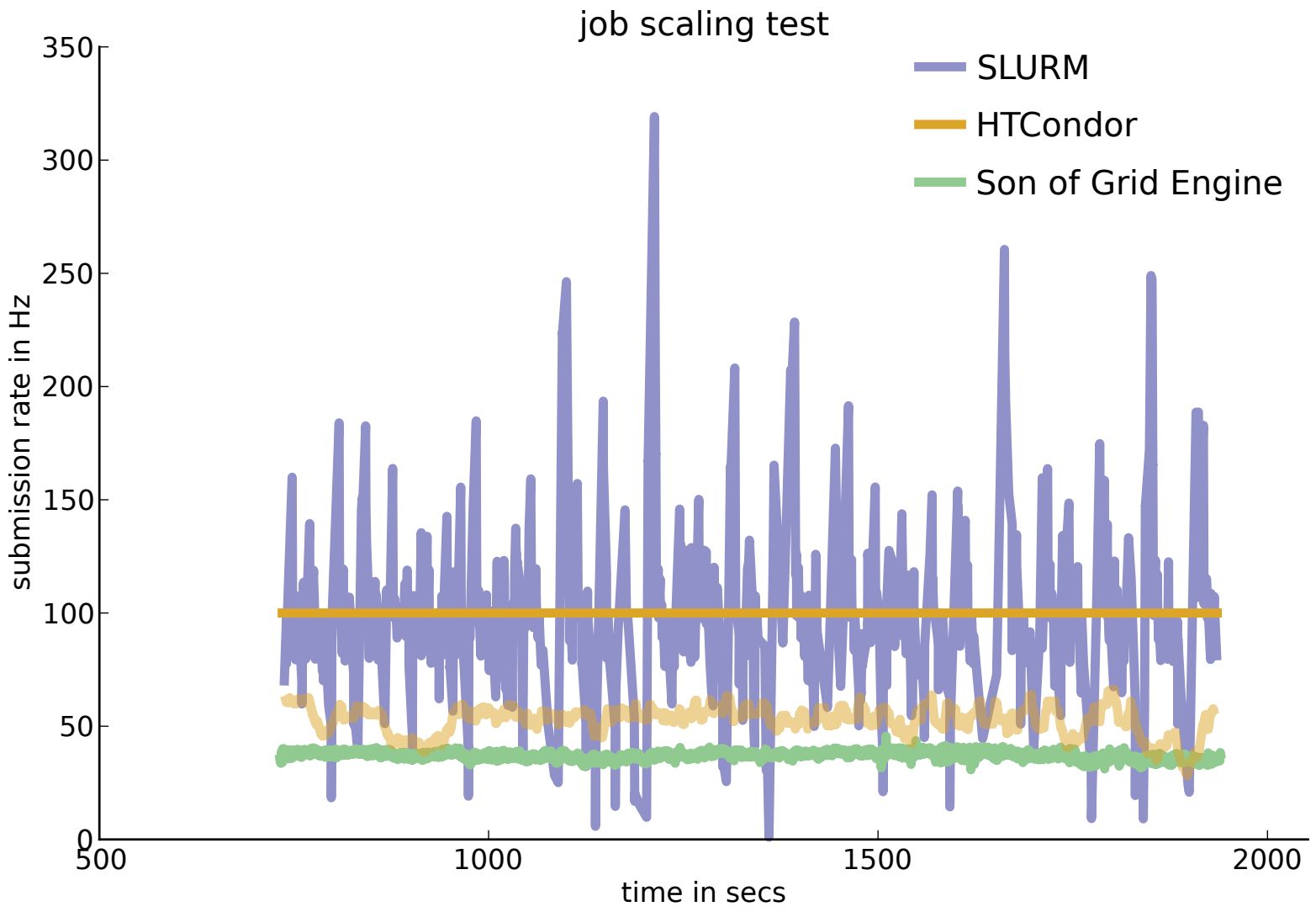
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Submission Rates



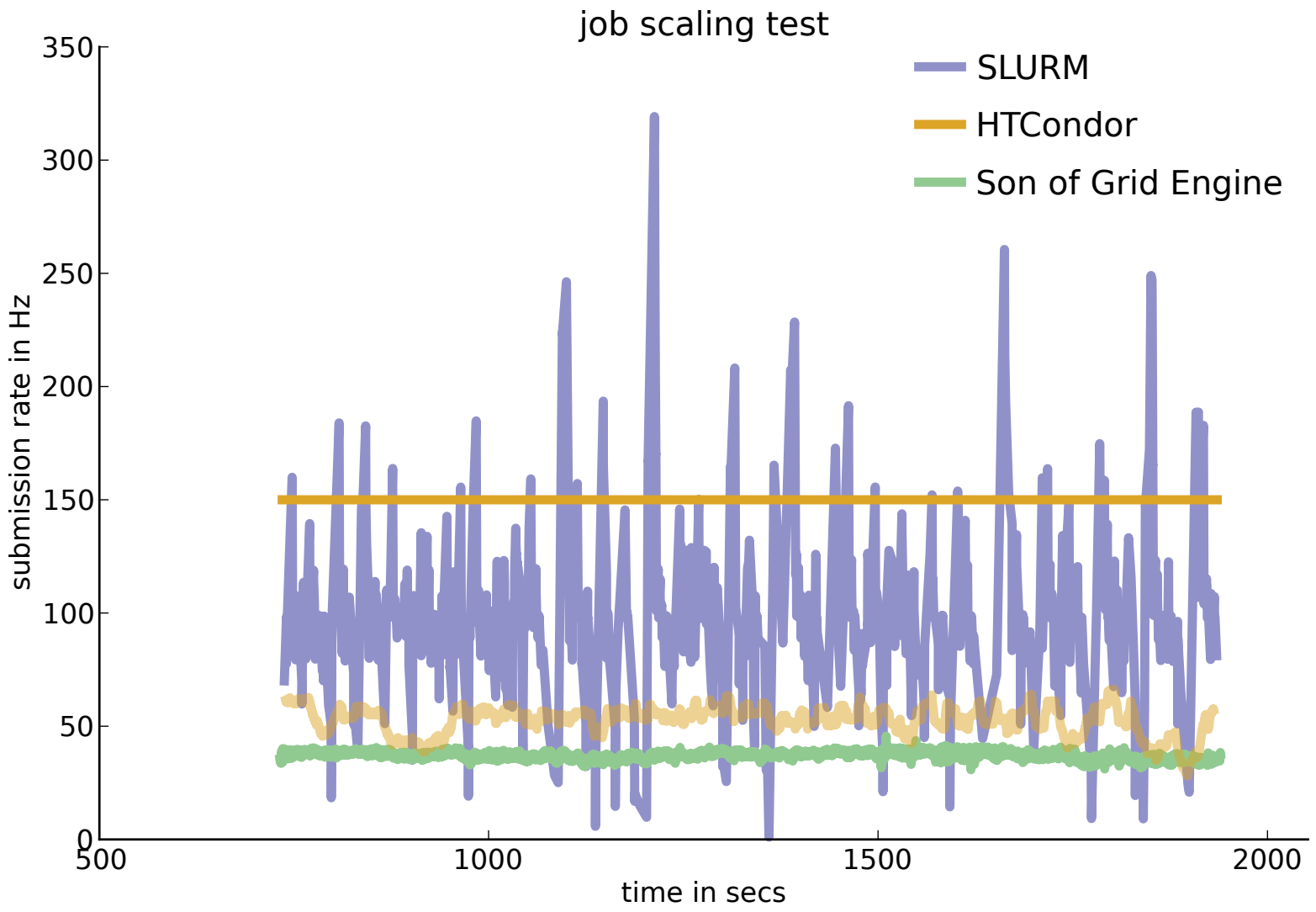
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Submission Rates



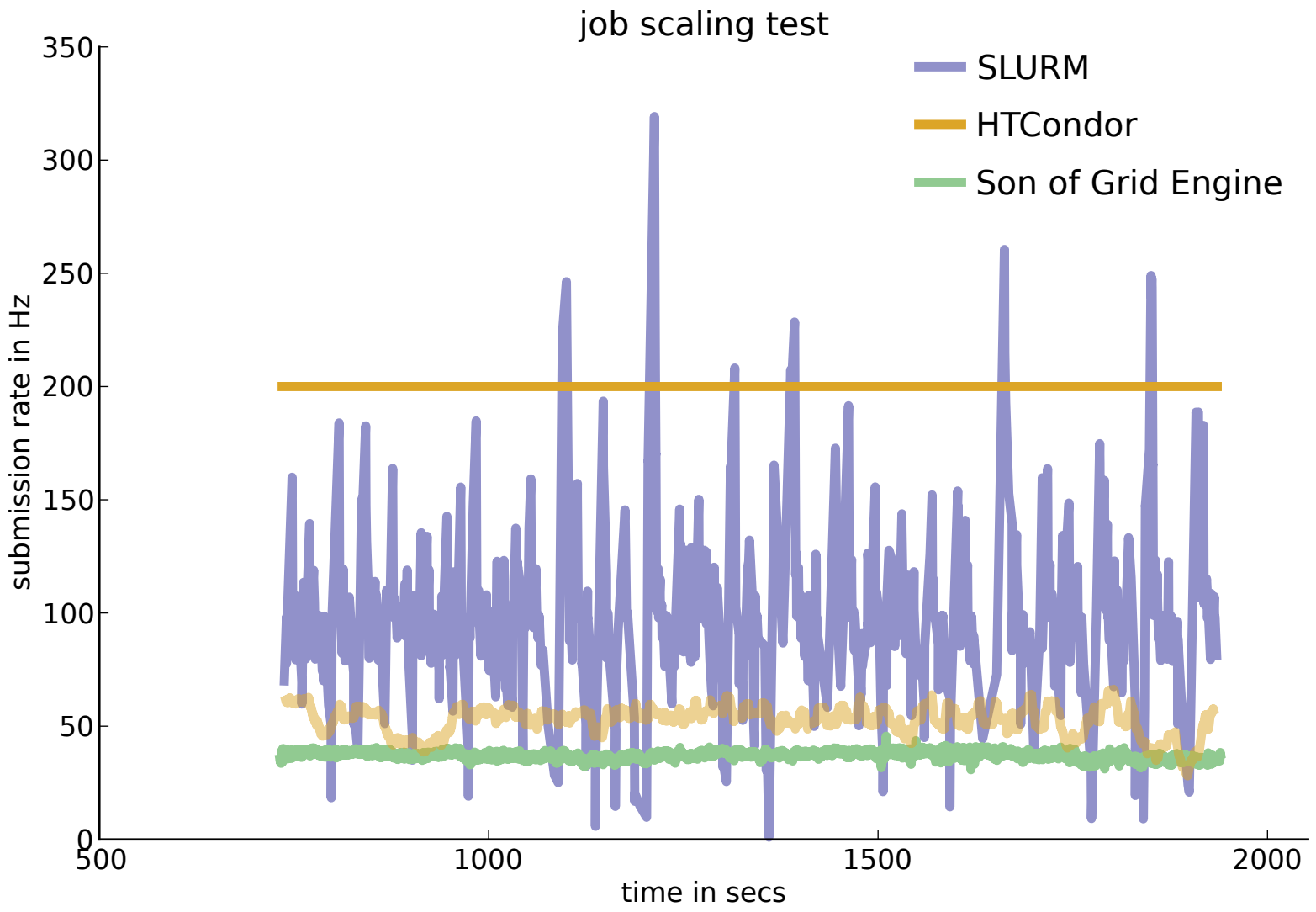
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Submission Rates



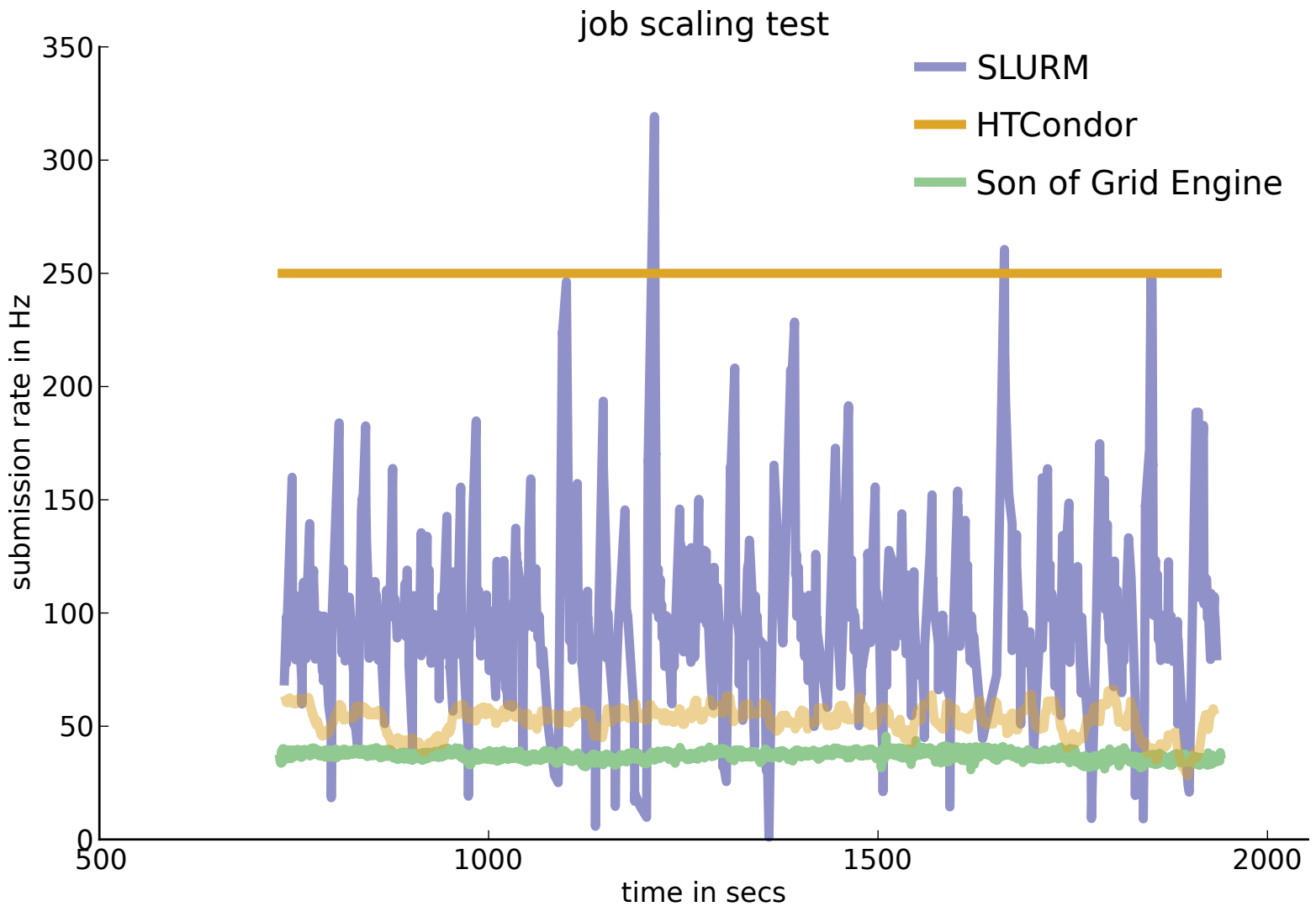
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Submission Rates



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Submission Rates



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Scalability

Requirements for measurement:

- Piggyback on existing resources
- Need to simulate many queries

Experience:

- Slow SLURM startup with too many hosts?
- HTCondor made to scale out
- CPU load on the master(s)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

General Impressions

	SLURM	HTCondor	Son of Grid Engine	Current LSF
Config (Puppet)	Struggle	Comfortable 1st time, updates harder	Needs shared filesystem	Complex
Maturity	Easy to freeze	Trustworthy	Rough around the edges	Trustworthy
Dynamic	Not really	Yes	Yes	Not really
Doc	Poor	Solid	Solid	Solid
Community support	Rather uninter-ested	Very enthusiastic	Enthusiastic	Commercial

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Required Features

- Grid support
- Kerberos/AFS
- Accounting
- Host normalisation
- Fairshare scheduling
- Support for commercial applications
- IPv6?

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Conclusions

Replacement candidates:

- SLURM feels too young
- HTCondor mature and promising
- Son of Grid Engine fast, a bit rough

What's next:

- Host scalability
- Query load
- Features

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Thanks!

Questions?

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it