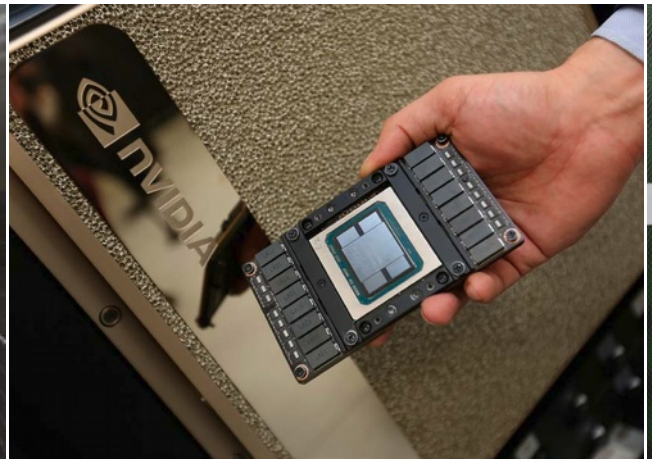
SciJava Interface for Parallelization



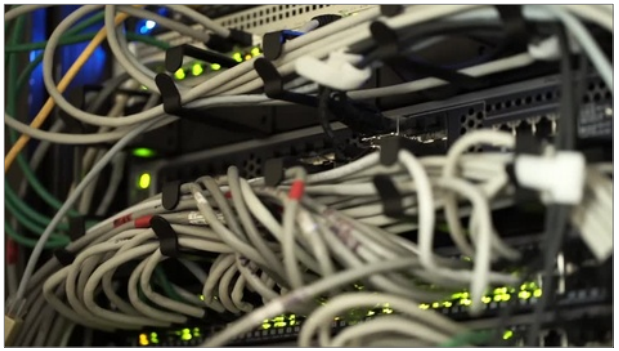
VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

HPC Cluster

- A set of inter-connected computers (nodes) with a shared storage
- Fault tolerance, parallel data processing, scalability, centralized management



Big Data in Life Sciences

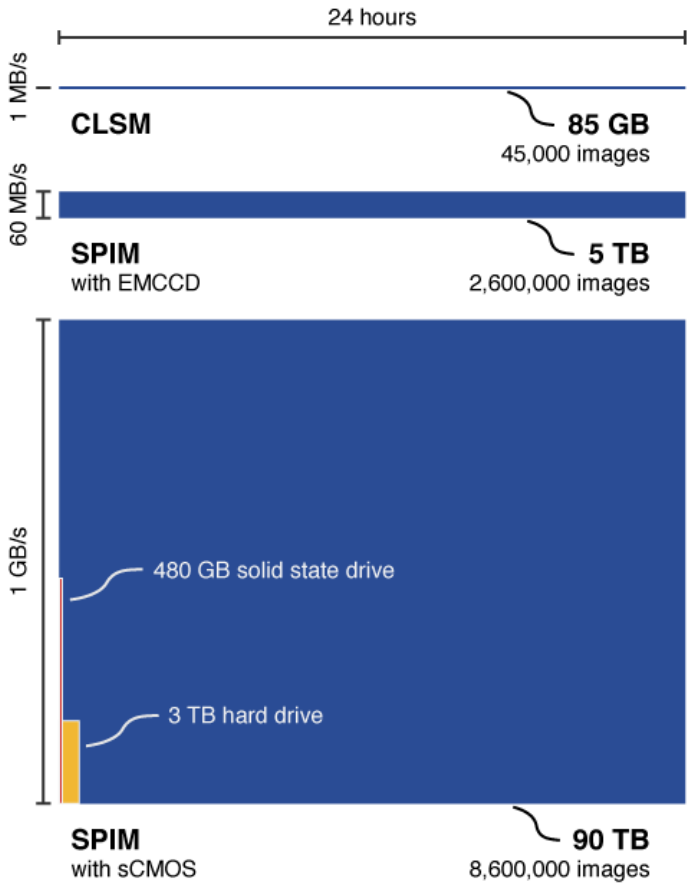


Image courtesy: Jan Huisken, Pavel Tomančák

SPIM Dataset (up to 90 TB)

Multi-view registration
Takes many hours

Multi-view fusion/deconvolution
Takes several days

To gain meaningful insights in a timely manner…
…using an HPC cluster seems like a good idea!

IT4Innovations National Supercomputing Center

Anselm (94 TFLOP/s)

209 compute nodes
- 16 cores, S. Bridge 2.4 GHz
- 64 GB RAM
- **23x NVIDIA Kepler K20**

Storage
- 500 TB

Network
- 40 Gb/s
- Infiniband QDR

Salomon (2011 TFLOP/s)

1,008 compute nodes
- 24 cores, Haswell-EP 2.5 GHz
- 128 GB RAM
- 864x Intel Xeon Phi 7120P

Storage
- 2,200 TB

Network
- 56 Gb/s
- Infiniband FDR

Nvidia DGX-2 (130 to 1920 TFLOP/s)

1 node
- 2 x Intel Xeon 8168
- Up to 1.5 TB RAM
- 512 GB HBM2 GPU m.
- **16 x NVIDIA Tesla V100 32GB HBM2**

Storage
- 30 TB NVMe (up to 60 TB total)

Network
- 8x Infiniband /100 GbE

1 TFLOP = 10^{12} FLOP (Floating-Point Operation)

IT4Innovations TOP500 06/19 National Supercomputing Center

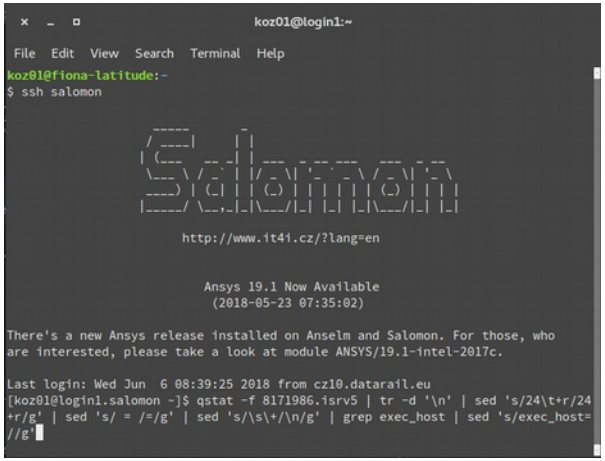
		Lenovo				
279	Internet Service A China	Inspur TS10000, Xeon Silver 4114 10C 2.2GHz, NVIDIA Tesla V100, Infiniband FDR Inspur	33,000	1,471.0	3,000.0	
280	Intel United States	**Endeavor** - Intel Cluster, Intel Xeon Gold 6148/Xeon Phi 7250F 68C 1.4GHz, Intel Omni-Path Intel	45,680	1,463.0	2,507.3	431
281	Service Provider T China	Lenovo HR650x, Xeon Gold 6133 20C 2.5GHz, 25G Ethernet Lenovo	33,600	1,462.3	2,688.0	
282	IT4Innovations National Supercomputing Center, VSB-Technical University of Ostrava Czech Republic	**Salomon** - SGI ICE X, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR, Intel Xeon Phi 7120P HPE	76,896	1,457.7	2,011.6	4,806
283	Government China	**N1** - ThinkSystem SD530, Xeon Gold 6140 18C 2.3GHz, NVIDIA Tesla V100, Infiniband EDR Lenovo	28,224	1,443.6	2,316.0	

Started as #85

https://www.top500.org

Cluster Access Challenges

- Command line interface
- **Bureaucracy**
- **Different job schedulers**
- **Data transfer speed limitation**
- Integration to the client application



Challenges in Depth – Job Schedulers

- Different HPC clusters use different schedulers
  - Portable Batch System – PBS
    - PBS Pro(fessional)
    - OpenPBS
    - TORQUE PBS
  - SLURM Workload Manager
  - Load Sharing Facility, Moab, LoadLeveler, Sun Grid Engine, …
  - And what if we decide to use Cloud instead?

Challenges in Depth – Job Schedulers

- Why is existence of more schedulers a problem?
- Different command line interface, different output, different environment variables/files

	PBS	SLURM
Start a job	qsub	sbatch
Query job status	qstat	squeue
Cancel a job	qdel	scancel
Job details	qstat -f	scontrol
...	...	...

Challenges in Depth – Data Transfer

- The clusters are typically not to be used for long-term data storage (others need to store their data and calculate as well)
- Bandwidth for transfers between compute nodes inside of the cluster is HUGE (> 10 Gb s^{-1} for a single TCP stream), but communication of compute node with outside is limited to 1 Gbs^{-1}
- The login nodes (which you access from outside) have connectivity exceeding 10 Gb s^{-1}
- So far so good… BUT

Challenges in Depth – Data Transfer

- Storage write speed is limited especially for huge number of small files (fast cache – Burst buffer – may be available on some system to speed it up)
- Your bandwidth may and WILL be limited by the distance to the cluster (Round-Trip-Time), network throughput and other factors (e.g. traffic shaping, NAT, IDS/IPS solution)
- Example:
  - Transferring files from DC in the same campus over SSH (scp) peaks at 109,35 MB s^{-1} (still bellow 1 Gb s^{-1})
  - Transfer of collection from MPI-CBG to Ostrava before 2019 Hackaton peaked at 400 Mb s^{-1} (~49,88 MB s^{-1}). Similarly, in June collections were sent at 361-420 Mb s^{-1}

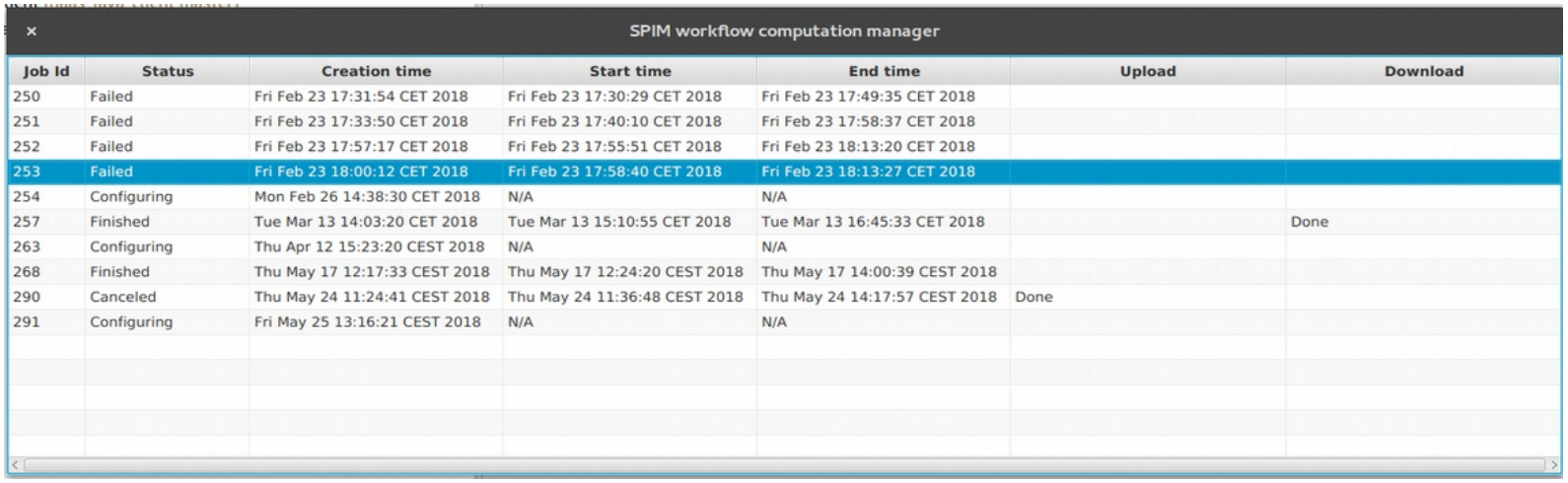
Mitigating Data Transfer Issues

We do not have time to transfer collections for several days if we need the results quickly.

- Solution 1: Do not use SSH, and ideally do not use TCP-based transfers
  - Issues: Security (especially data confidentiality), UDP may be intentionally delayed/dropped by carriers
- Solution 2: Parallelize the data transfers – this leads to significant speed up to 8 Gb s^{-1} for 10 concurrent transfers (but most SSH servers will not allow you more)
- Solution 3: Upload the data as they are being created/collected

Plugin-specific Solution

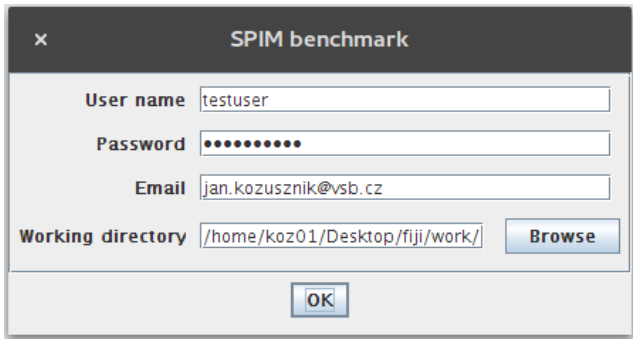
HEAppE (HPCaaS) middleware was successfully integrated into a plugin performing the Snakemake pipeline[1] for SPIM data processing

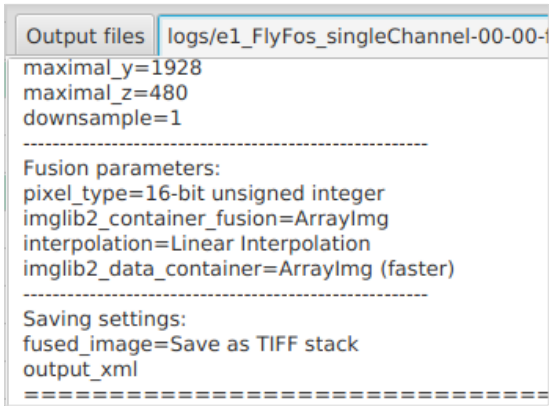


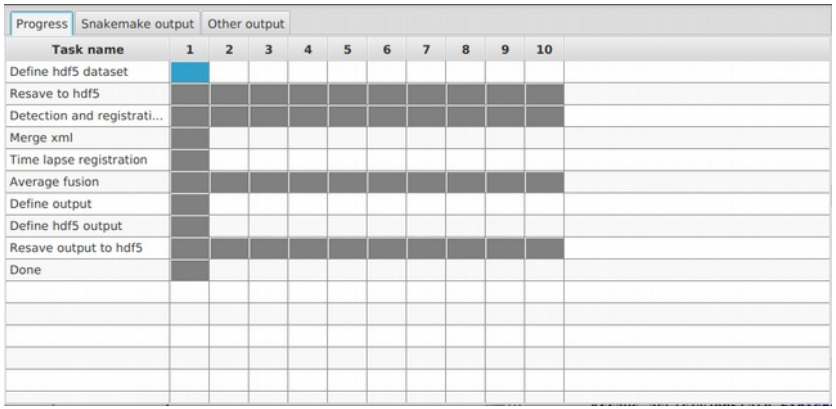
Job Id	Status	Creation time	Start time	End time	Upload	Download
250	Failed	Fri Feb 23 17:31:54 CET 2018	Fri Feb 23 17:30:29 CET 2018	Fri Feb 23 17:49:35 CET 2018		
251	Failed	Fri Feb 23 17:33:50 CET 2018	Fri Feb 23 17:40:10 CET 2018	Fri Feb 23 17:58:37 CET 2018		
252	Failed	Fri Feb 23 17:57:17 CET 2018	Fri Feb 23 17:55:51 CET 2018	Fri Feb 23 18:13:20 CET 2018		
253	Failed	Fri Feb 23 18:00:12 CET 2018	Fri Feb 23 17:58:40 CET 2018	Fri Feb 23 18:13:27 CET 2018		
254	Configuring	Mon Feb 26 14:38:30 CET 2018	N/A	N/A		
257	Finished	Tue Mar 13 14:03:20 CET 2018	Tue Mar 13 15:10:55 CET 2018	Tue Mar 13 16:45:33 CET 2018		Done
263	Configuring	Thu Apr 12 15:23:20 CEST 2018	N/A	N/A		
268	Finished	Thu May 17 12:17:33 CEST 2018	Thu May 17 12:24:20 CEST 2018	Thu May 17 14:00:39 CEST 2018		
290	Canceled	Thu May 24 11:24:41 CEST 2018	Thu May 24 11:36:48 CEST 2018	Thu May 24 14:17:57 CEST 2018	Done	
291	Configuring	Fri May 25 13:16:21 CEST 2018	N/A	N/A		

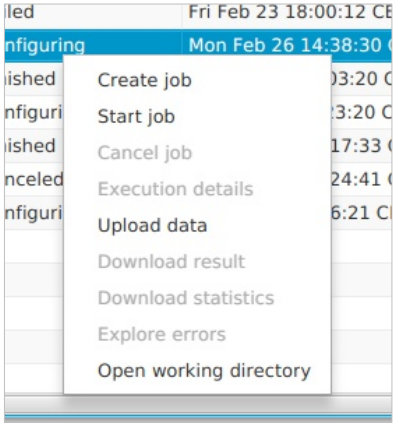
[1] Schmied, C., Steinbach, P., Pietzsch, T., Preibisch, S., Tomancak, P.: An automated workflow for parallel processing of large multiview spim recordings. Bioinformatics 32(7), 1112–1114 (2016)

Plugin-specific Solution









We want an ImageJ-wide Approach

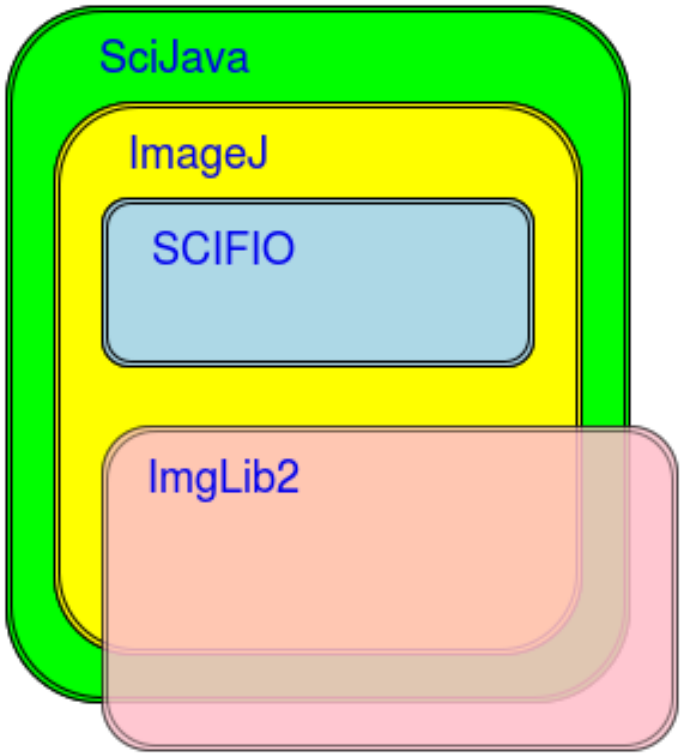
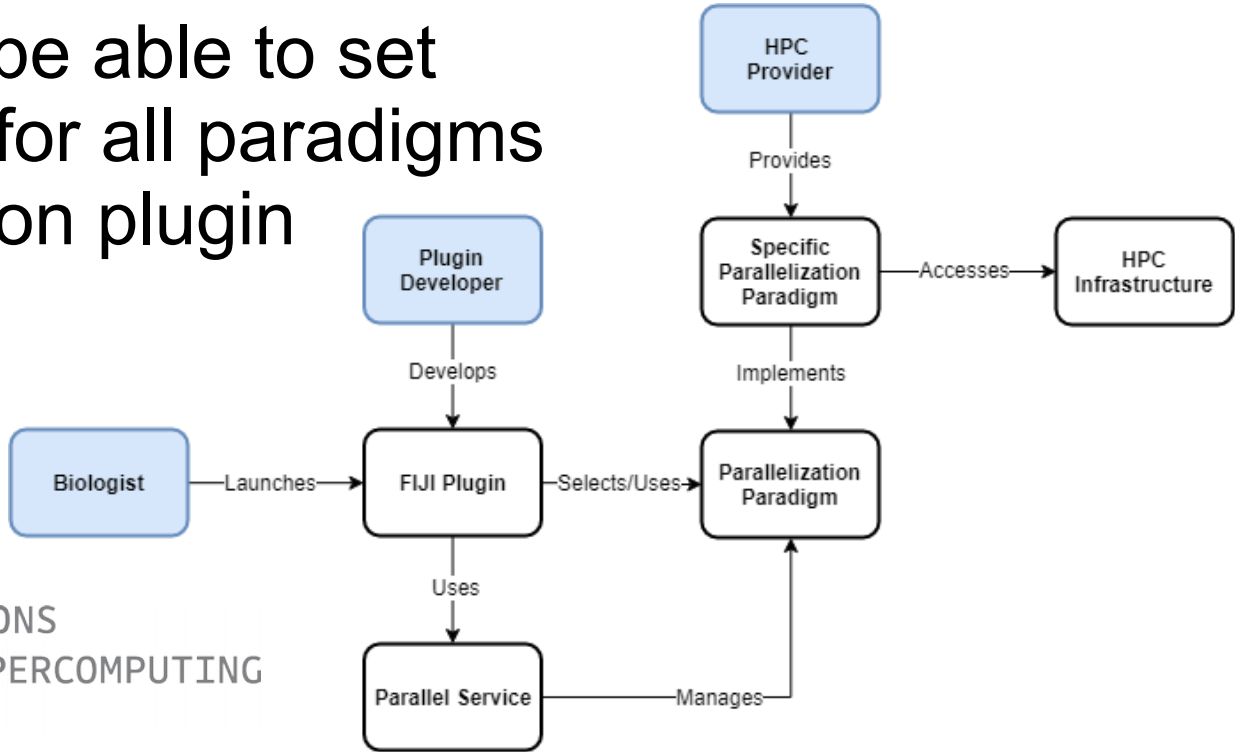


Image courtesy: https://imagej.net/Architecture

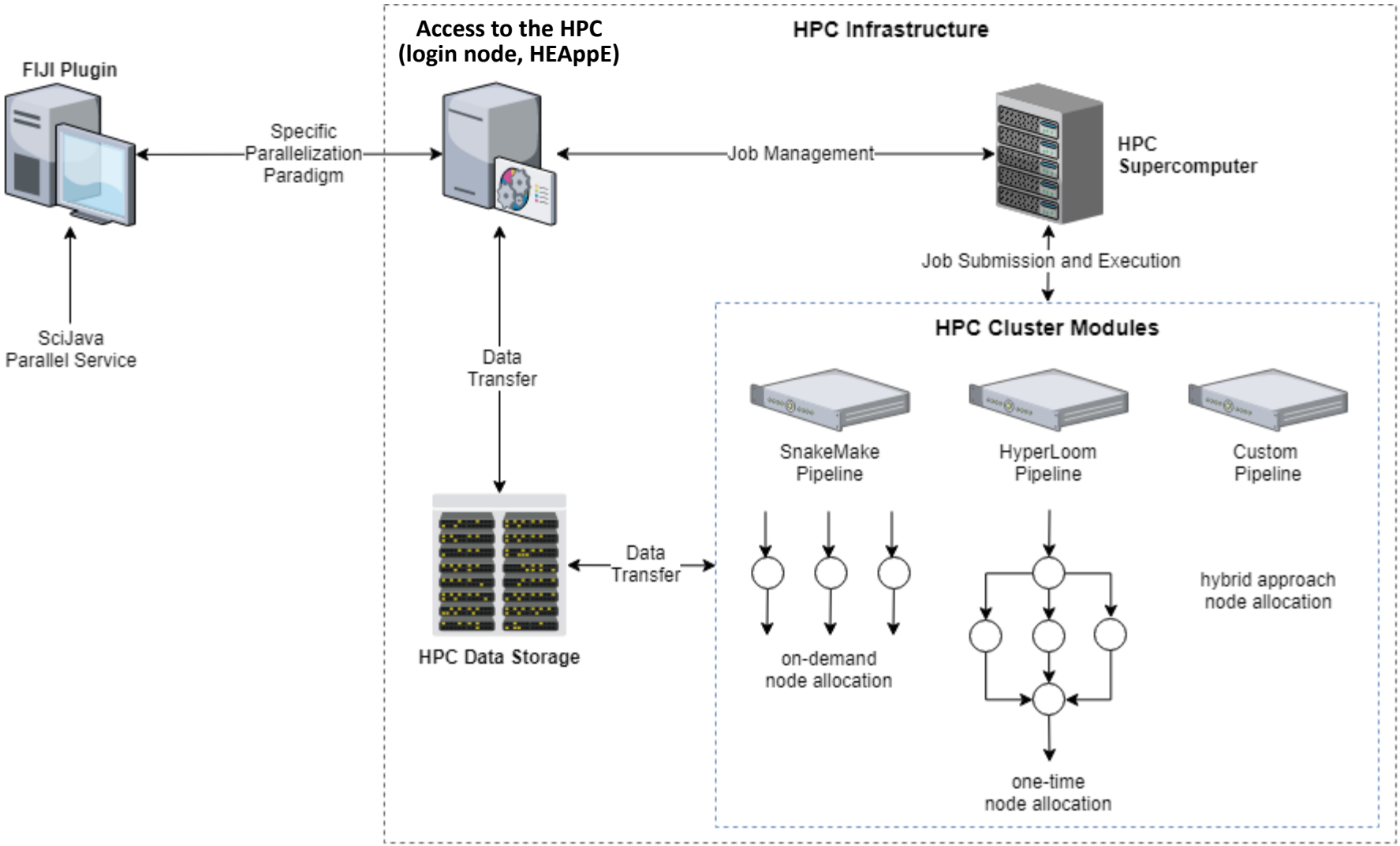
- ImageJ is extremely extensible
- New plugins are constantly being developed
- Parallelization support shall be integrated into a core library, such as SciJava

So let's make the SciJava Parallel…

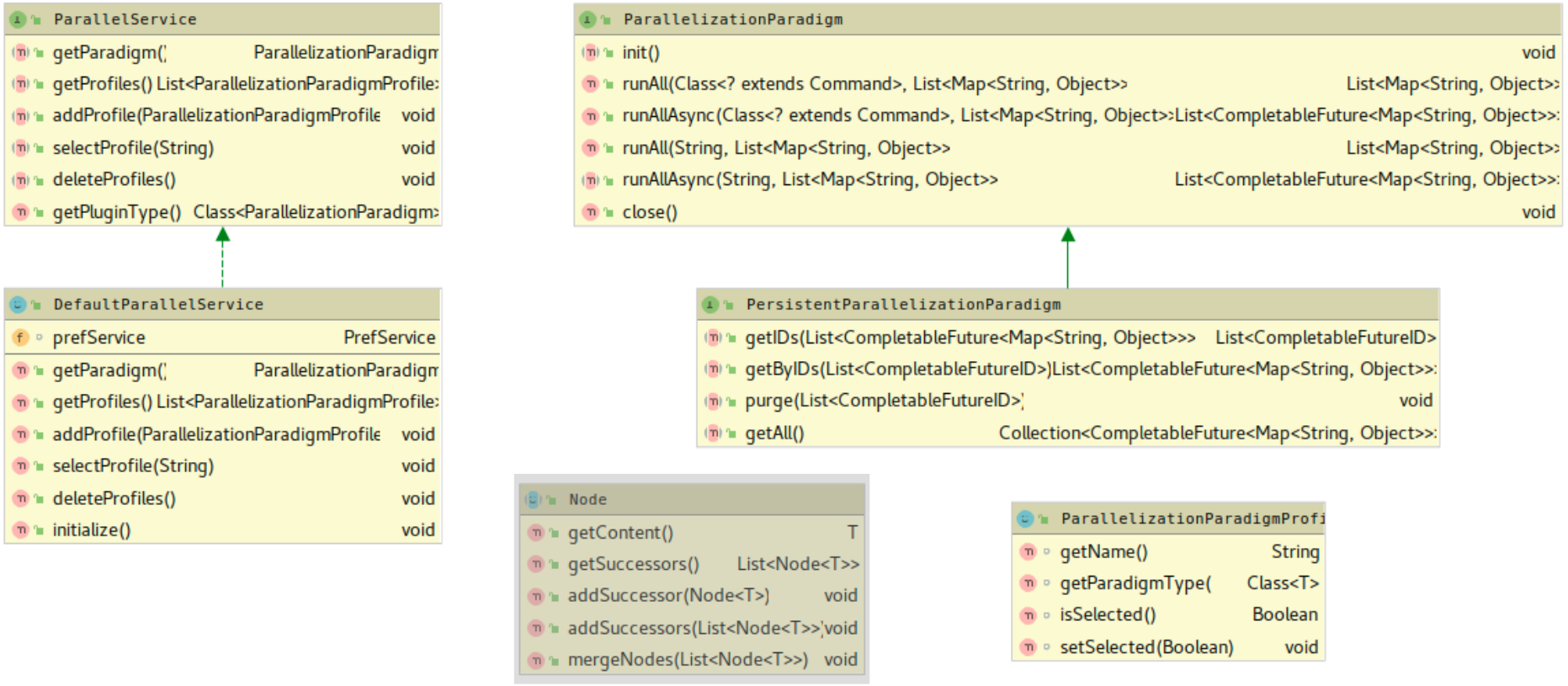
- SciJava singleton service managing *parallelization paradigms*
- Paradigms implementing the defined API will be available at update sites
- End users will be able to set up credentials for all paradigms in a configuration plugin



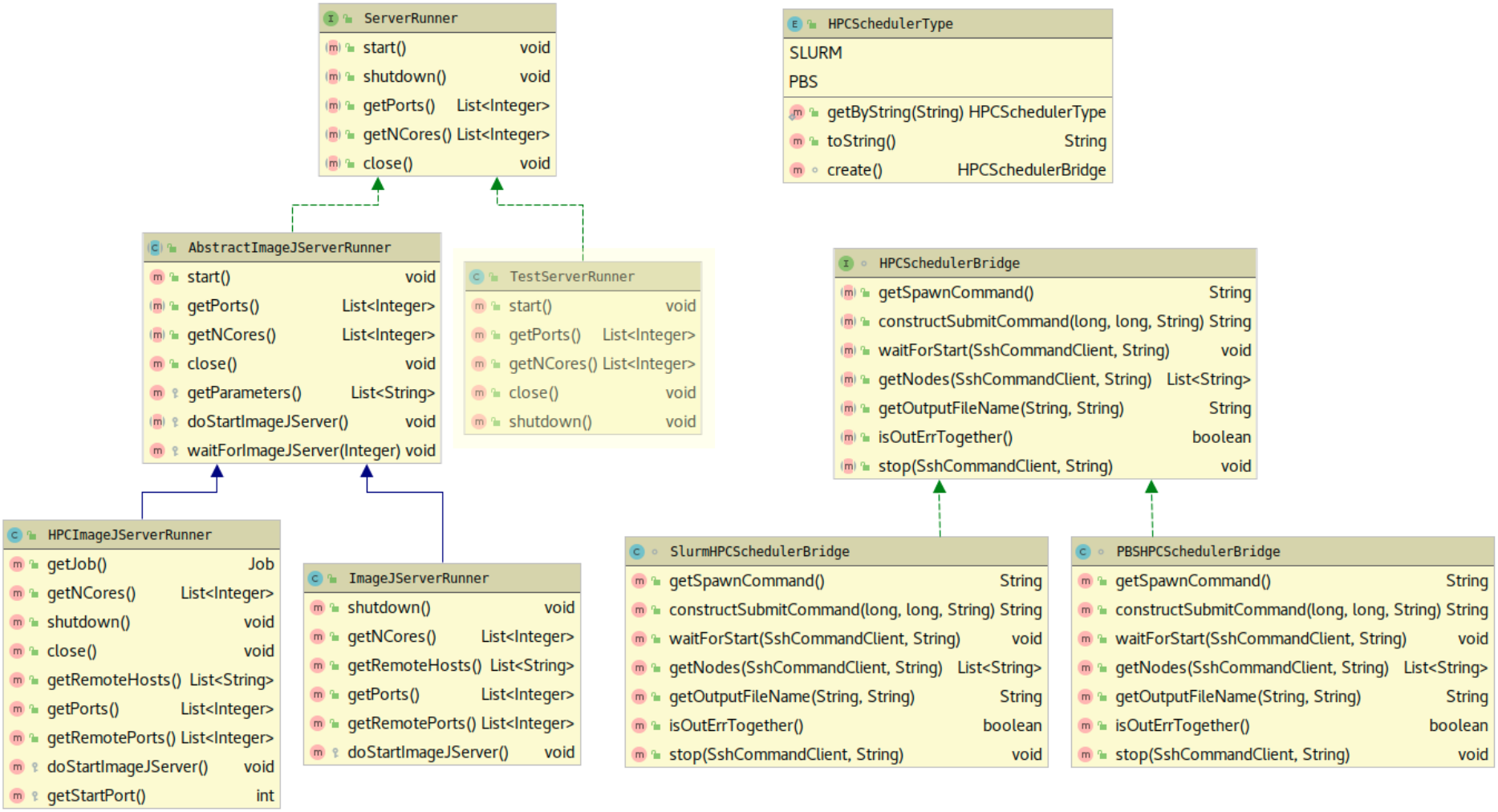
How does it look in HPC environment?



ParallelService



Under the Hood – Running on HPC



ParallelService

- Plugin developers can use available paradigms for remote computations

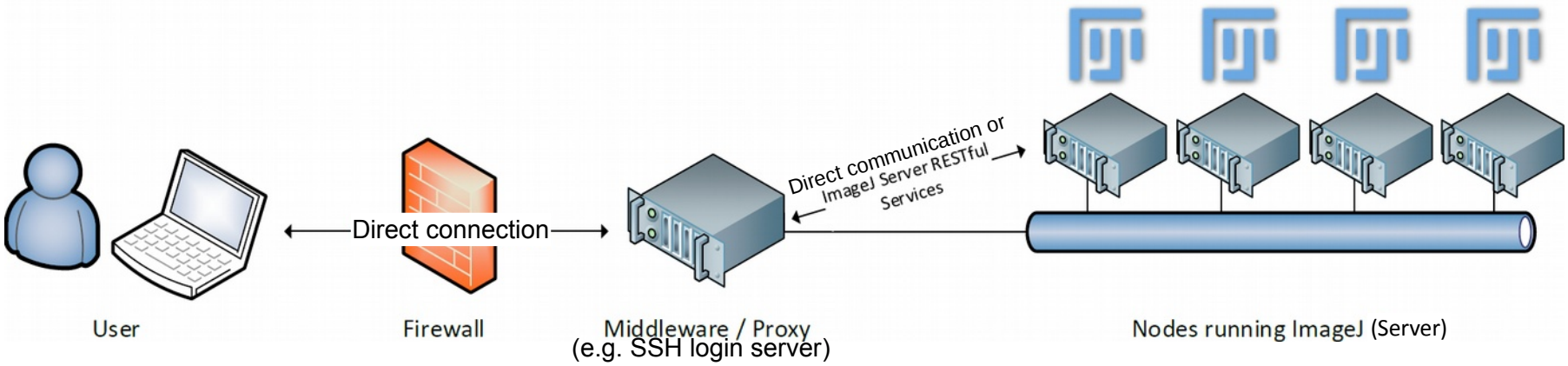
```
parallelService.selectProfile(...);
try (ParallelizationParadigm paradigm = parallelService.getParadigm()) {
    paradigm.init();
    List<Map<String, Object>> results = paradigm.runAll(Command.class, p_list);
    results.forEach(...);
}
```


EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

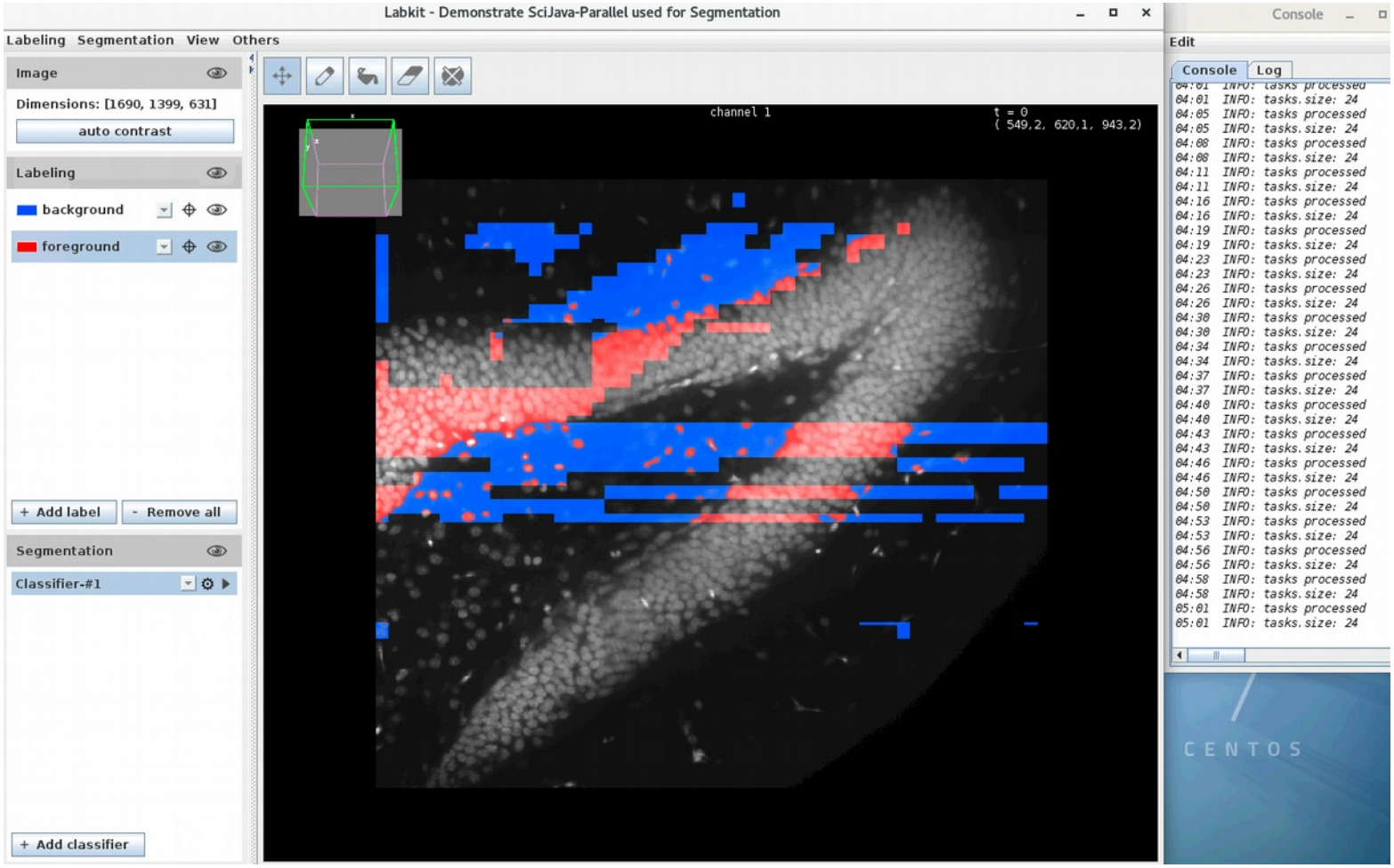
Parallelization Paradigm Example

- Experimental paradigm and ImageJ Server instances running on nodes
- Live demo will follow (sorry, no credentials)…



Parallel LabKit Use Case

Run the LabKit plugin automatic segmentation in parallel on multiple nodes of a cluster



Discussion

- Other use cases
- Requirements
- Some stories from our backlog
  - Local and remote folder syncing
  - Cost-efficiency assessment tool

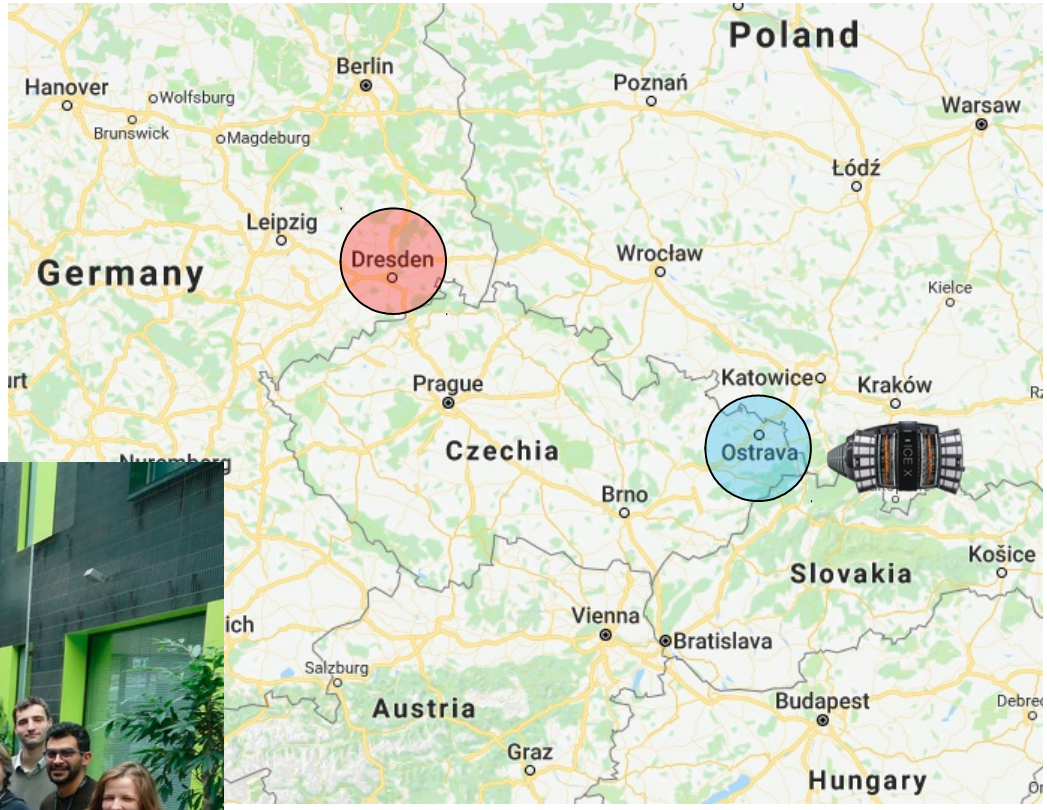
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Hackathon in Ostrava

Preliminary date
TBA

Accommodation
Campus hotel



Airports
Ostrava (20 mins)
Katowice (1.5 hours)
Krakow (2 hours)
Prague (3.5 hours)
Vienna (3.5 hours)

Thank you very much for your attention

This work was supported by the European Regional Development Fund in the IT4Innovations national supercomputing center – path to exascale project,
project number CZ.02.1.01/0.0/0.0/16 013/0001791 within the Operational Programme Research, Development and Education.