GHENT UNIVERSITY

HPC-UGent pilot follow up meeting

May 16th 2018

http://users.ugent.be/~kehoste/hpcugent-pilot-followup-20180516.pdf

hpc@ugent.be

http://ugent.be/hpc

`skitty` & `victini`

- new HPC-UGent Tier-2 clusters
- `module swap cluster/.skitty`
  - replacement for delcatty (to be retired soon)
  - **72** nodes, each with 2x18-core Intel Skylake + **192GB** RAM, **EDR Infiniband**

- `module swap cluster/.victini`
  - replacement for raichu (already retired)
  - **96** nodes, each with 2x18-core Intel Skylake + **96GB** RAM, **10Gb Ethernet**

- *only accessible to pilot users for now*

GHENT UNIVERSITY

Differences with existing Tier-2

- **SLURM** as resource manager (instead of Torque/PBS)
  - Torque/PBS is not sustainable due to support issues
  - wrappers are in place to make this switch **transparent** to users
  - `qsub`, `qstat`, `qalter`, `qdel` commands should still work
  - `#PBS` header lines in job scripts should still work
  - `$PBS_*` environment variables should still be defined
- software installations: *only with 2018a toolchains* (& newer)

GHENT UNIVERSITY

Timeline

- *March 14th:* pilot kickoff, skitty available for pilot users
- March 26th: victini also available for pilot users
- *May 7th:* additional pilot users added
- ***May 16th (today): pilot follow-up meeting***
- *summer 2018:* skitty & victini go into production
- *Fall 2018:* switch default cluster to victini + retire delcatty cluster
- *mid 2019:* switch to SLURM on all Tier-2 clusters

GHENT UNIVERSITY

Software installations

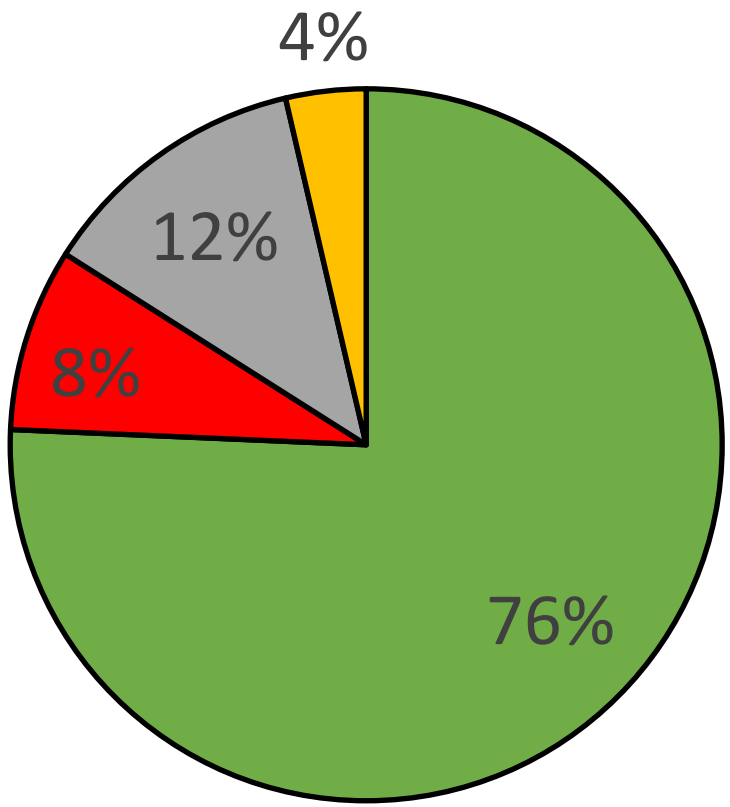
	intel/2018a	foss/2018a	binary
ABAQUS	-	-	**OK**
AmberTools	**OK**	*TODO (?)*	-
CP2K	**OK**	**OK**	-
dask	**OK**	*TODO*	-
FLUENT	-	-	**OK**
Gaussian	**OK**	*(skip)*	-
ISCE	*(skip)*	**OK (user)**	-
MATLAB	-	-	**OK**
molmod	**OK**	**OK**	-
OpenFOAM	**OK**	***TODO***	-
phonopy	**OK**	**OK**	-
Python	**OK**	**OK**	-
R	**OK**	***TODO***	-
StaMPS	*(skip)*	**OK (user)**	-
VASP	**OK**	*(skip?)*	-
yaff	**OK**	**OK**	-

GHENT UNIVERSITY

Total utilisation

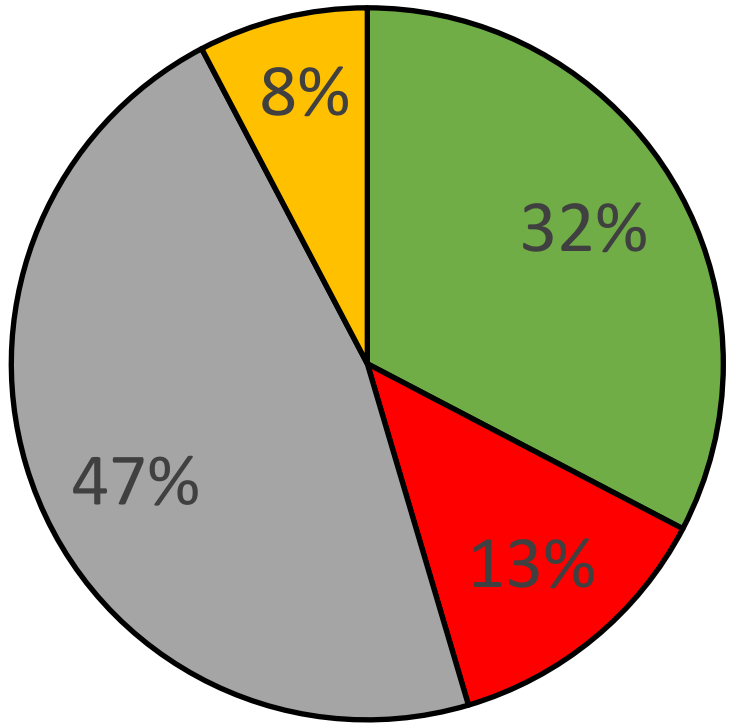
skitty - utilisation

(May 16th 2018 - ~2 months of pilot)



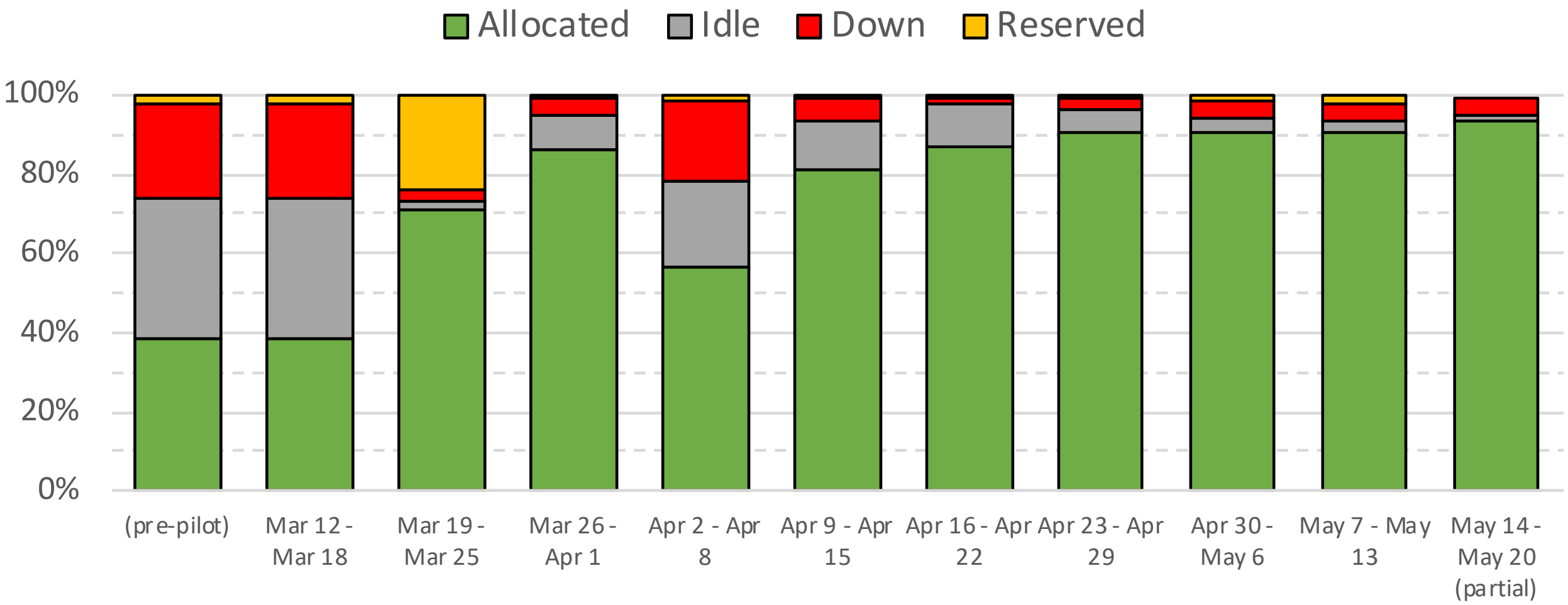
victini - utilisation

(May 16th 2018 - ~2 months of pilot)



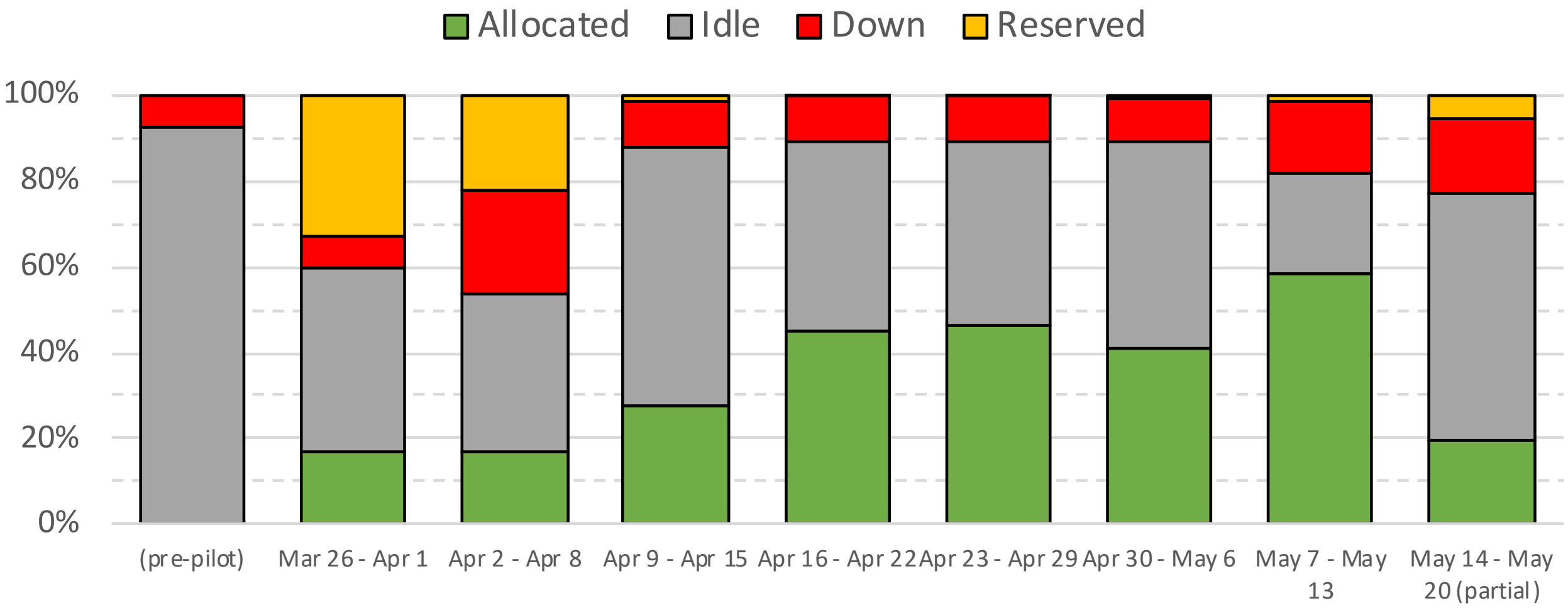
GHENT UNIVERSITY

Total utilisation (skitty)



GHENT UNIVERSITY

Total utilisation (victini)

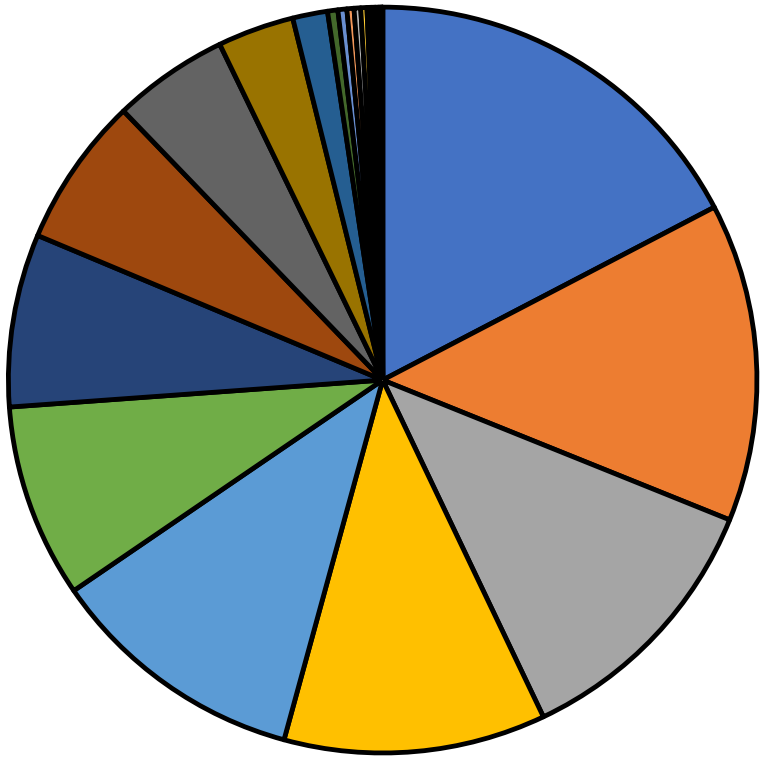


GHENT UNIVERSITY

Usage per user (anom.)

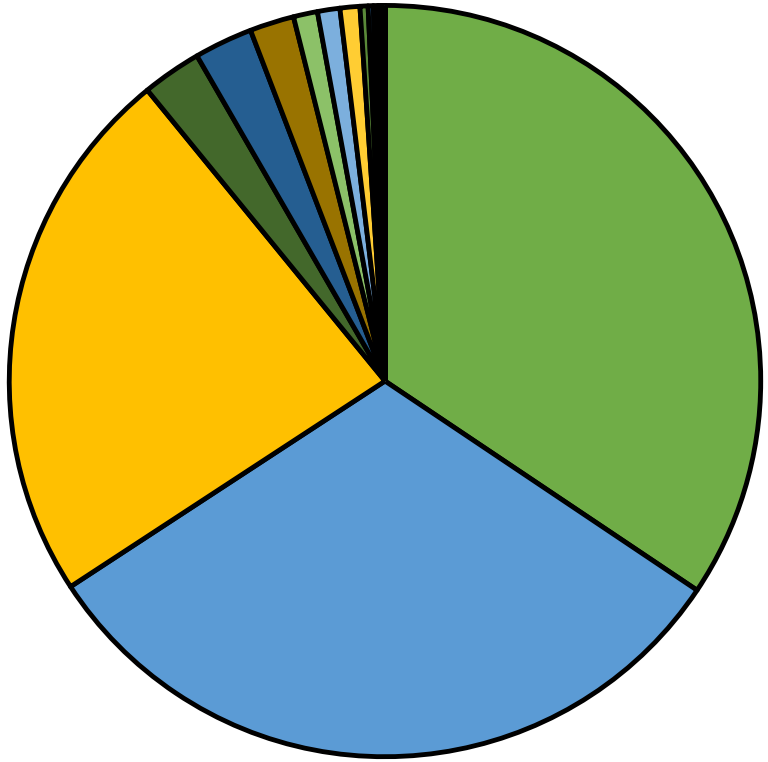
skitty - usage per user

(May 16th 2018 - ~2 months of pilot)



victini - usage per user

(May 16th 2018 - ~2 months of pilot)



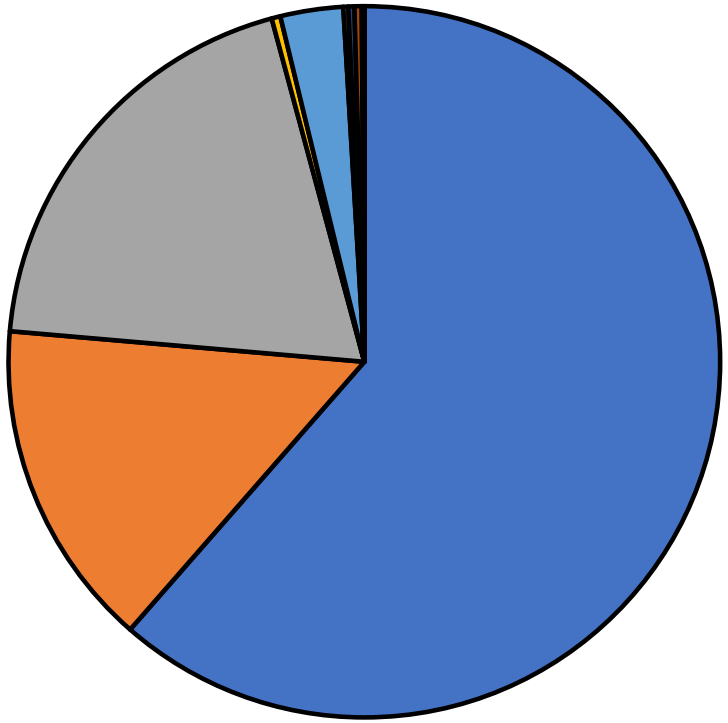
*(note: colors are *not* synced between both graphs)*

GHENT UNIVERSITY

Usage per VO (anom.)

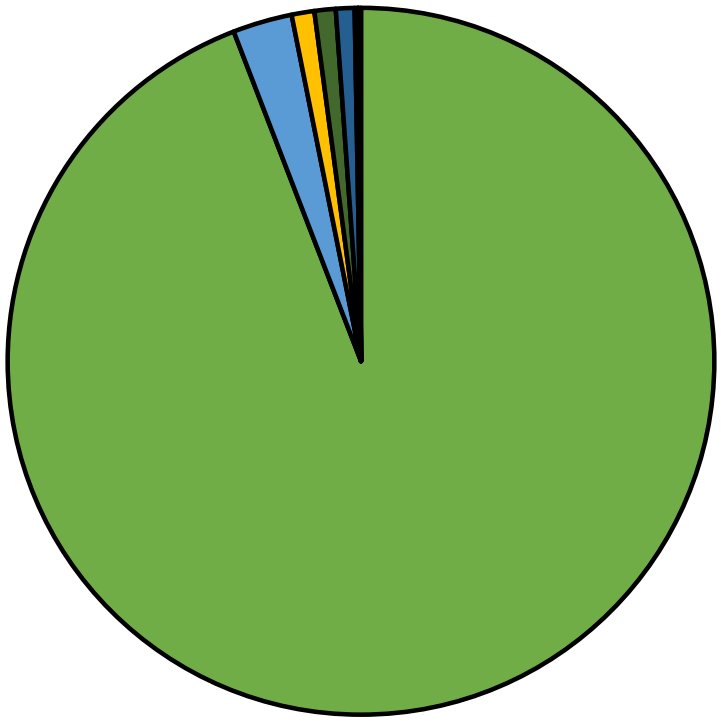
skitty - usage per VO

(May 16th 2018 - ~2 months of pilot)



victini - usage per VO

(May 16th 2018 - ~2 months of pilot)



*(note: colors are *not* synced between both graphs)*

GHENT UNIVERSITY

Solved issues since kickoff

- several issues with SLURM wrappers have been fixed
  - qsub, qstat, qdel should work as expected
  - `#PBS ...` and `${PBS_*}` in job scripts work as intended
- mympirun is now stable on top of SLURM (v4.1.1)
- faulty default memory limits, resulting in submitted jobs never starting
- node health check scripts are in place
- `/tmp` is now cleaned up automatically
- better job isolation (private `/tmp` & `/dev/shm` namespace per job, ...)
- (and lots more behind the scenes...)

GHENT UNIVERSITY

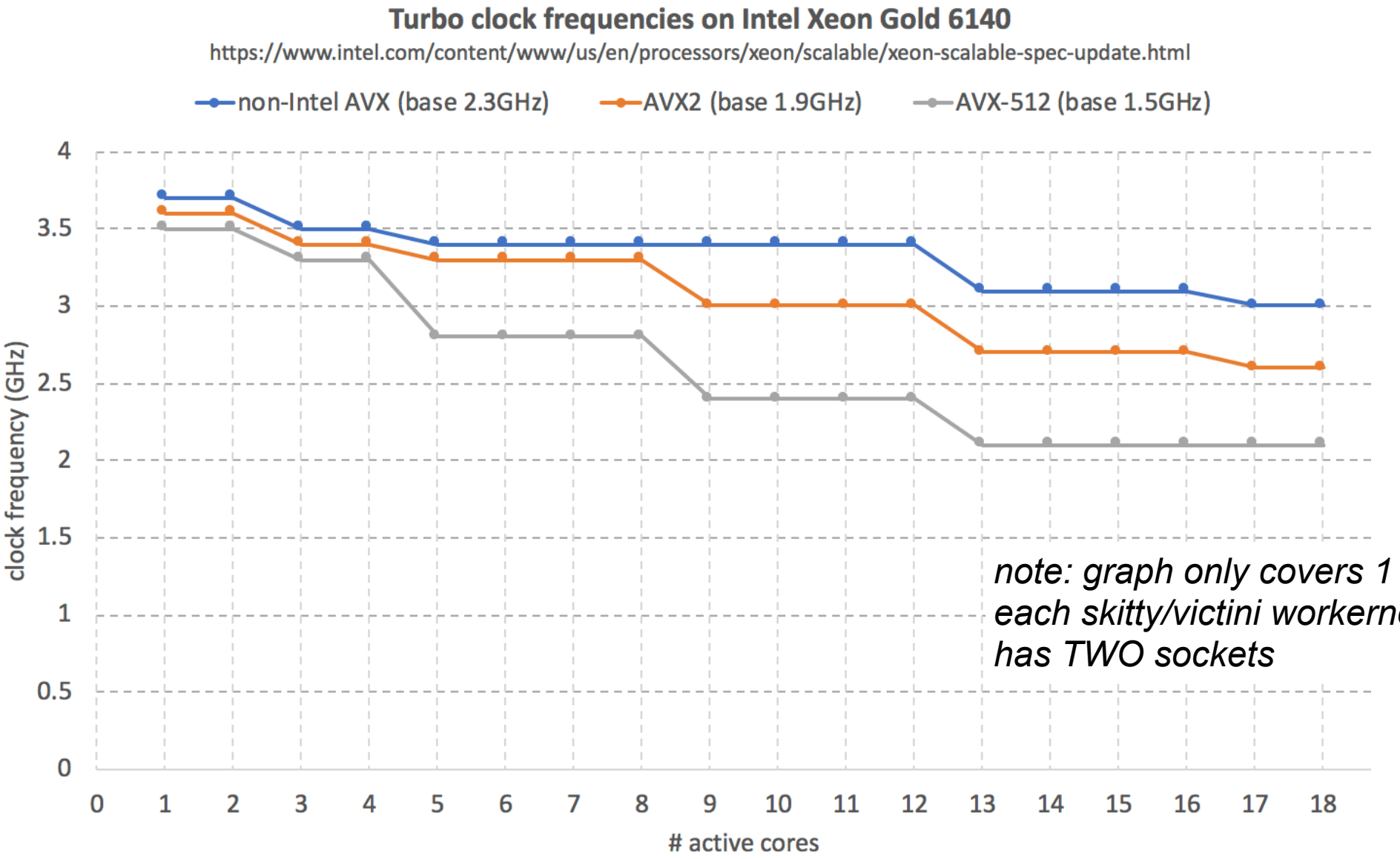
Known issues

- several down nodes in both skitty & victini
- `pbsmon` is not available yet (but `sinfo` can be useful too)
- `qalter` wrapper for SLURM needs work
- MPI hangs in interactive jobs
- ~~`$VSC_SCRATCH` and `$VSC_DATA` can't be used in 'old' VSC accounts (< vsc40900)~~

  ~~(use `$VSC_SCRATCH_VO_USER` and `$VSC_DATA_VO_USER` instead)~~
- custom software installations for `victini`: OpenMPI, GDAL
- some software-specific problems with VASP & CP2K
- performance benefit of optimising for AVX-512 is unclear...

GHENT UNIVERSITY

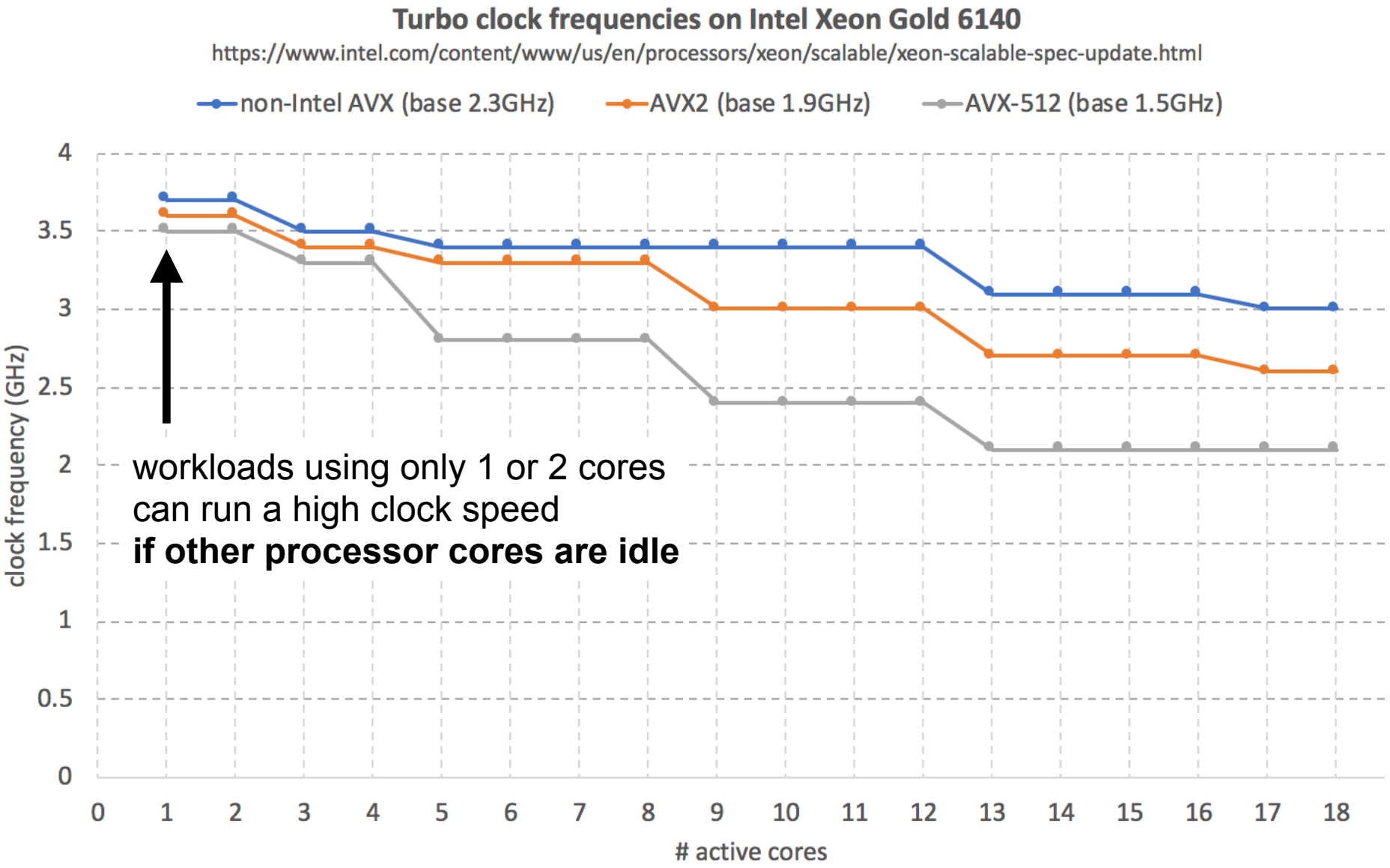
Intel Skylake: AVX-512 vs turbo



note: graph only covers 1 socket, each skitty/victini workernode has TWO sockets

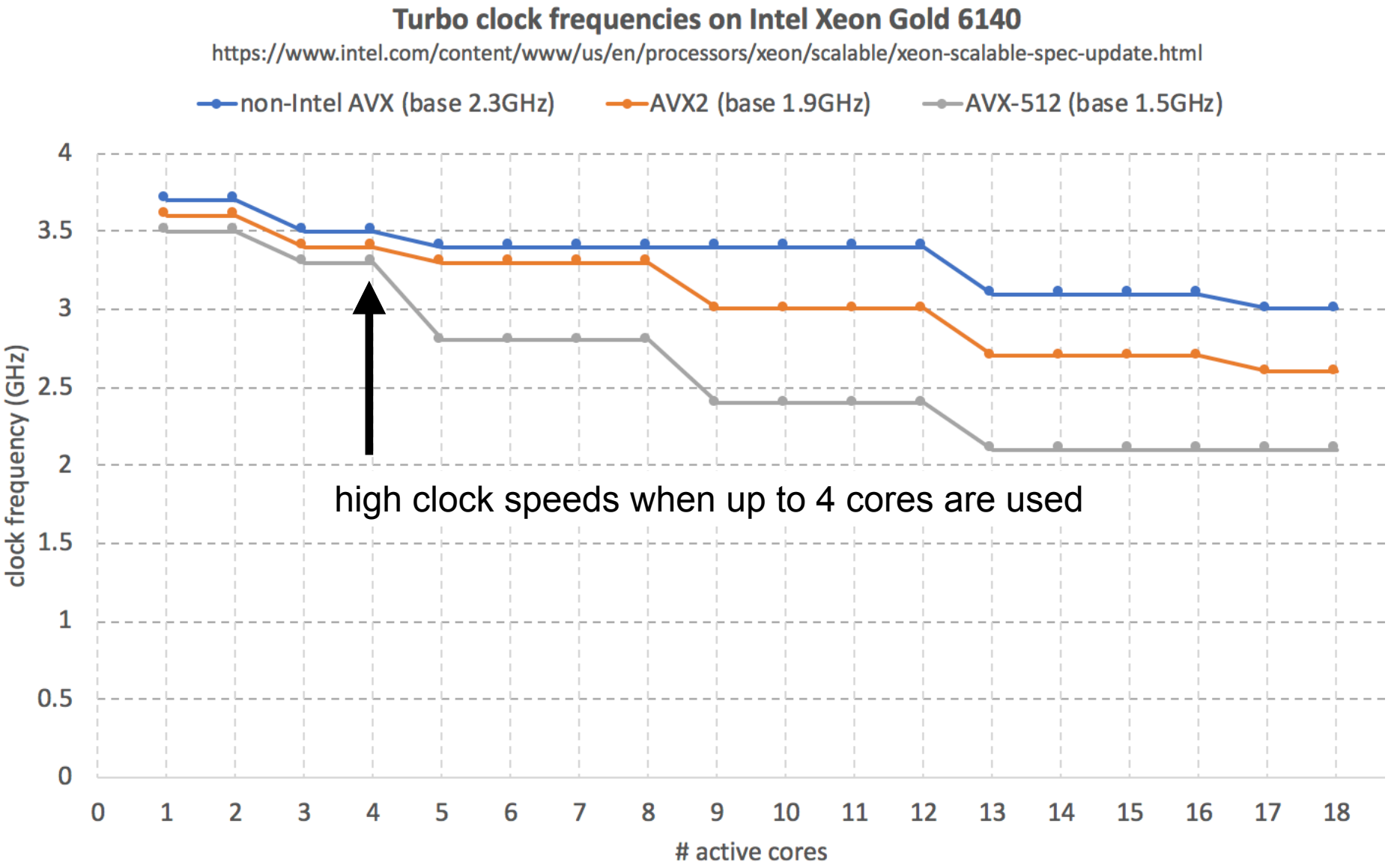
GHENT UNIVERSITY

Intel Skylake: AVX-512 vs turbo



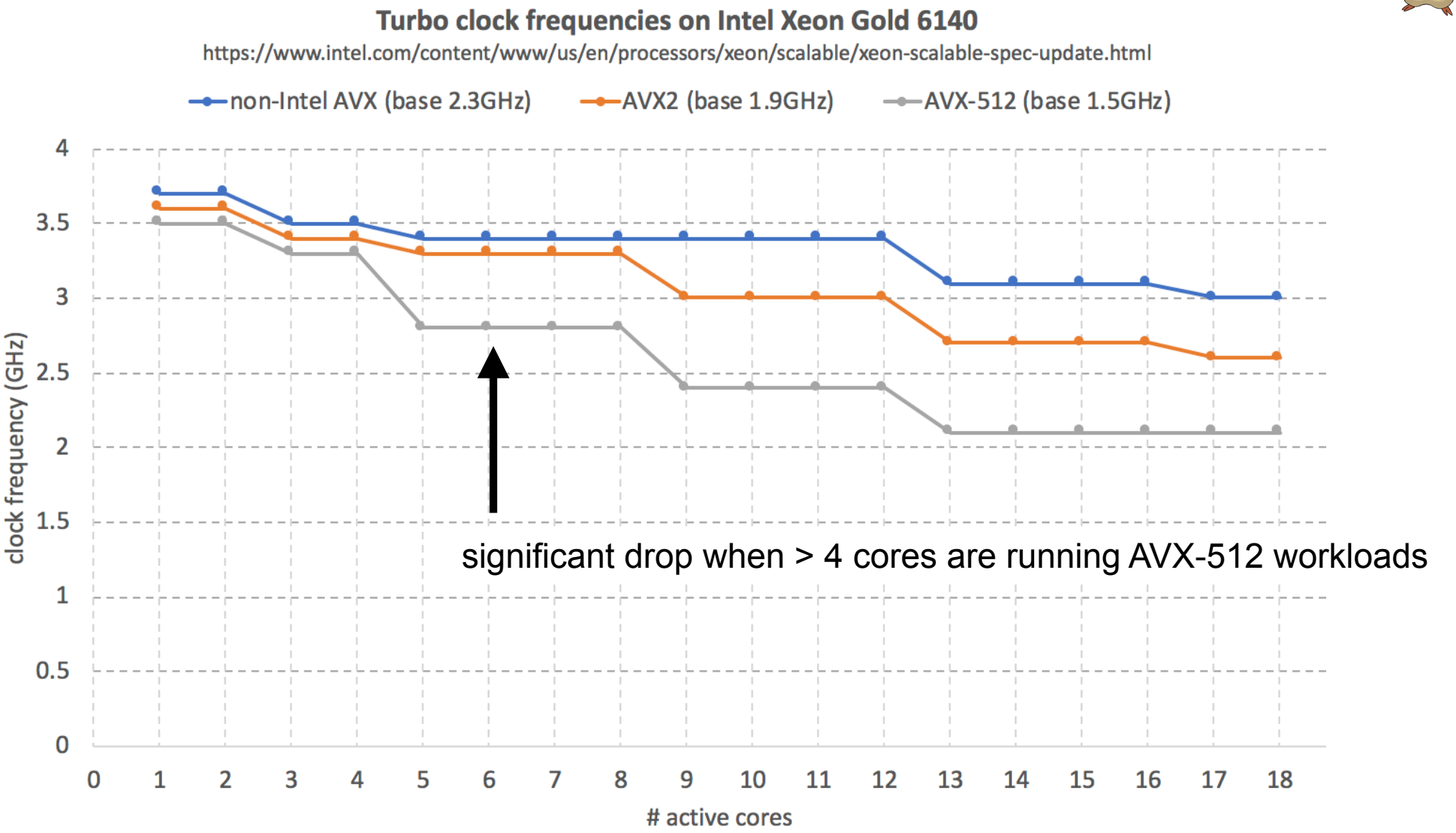
GHENT UNIVERSITY

Intel Skylake: AVX-512 vs turbo



GHENT UNIVERSITY

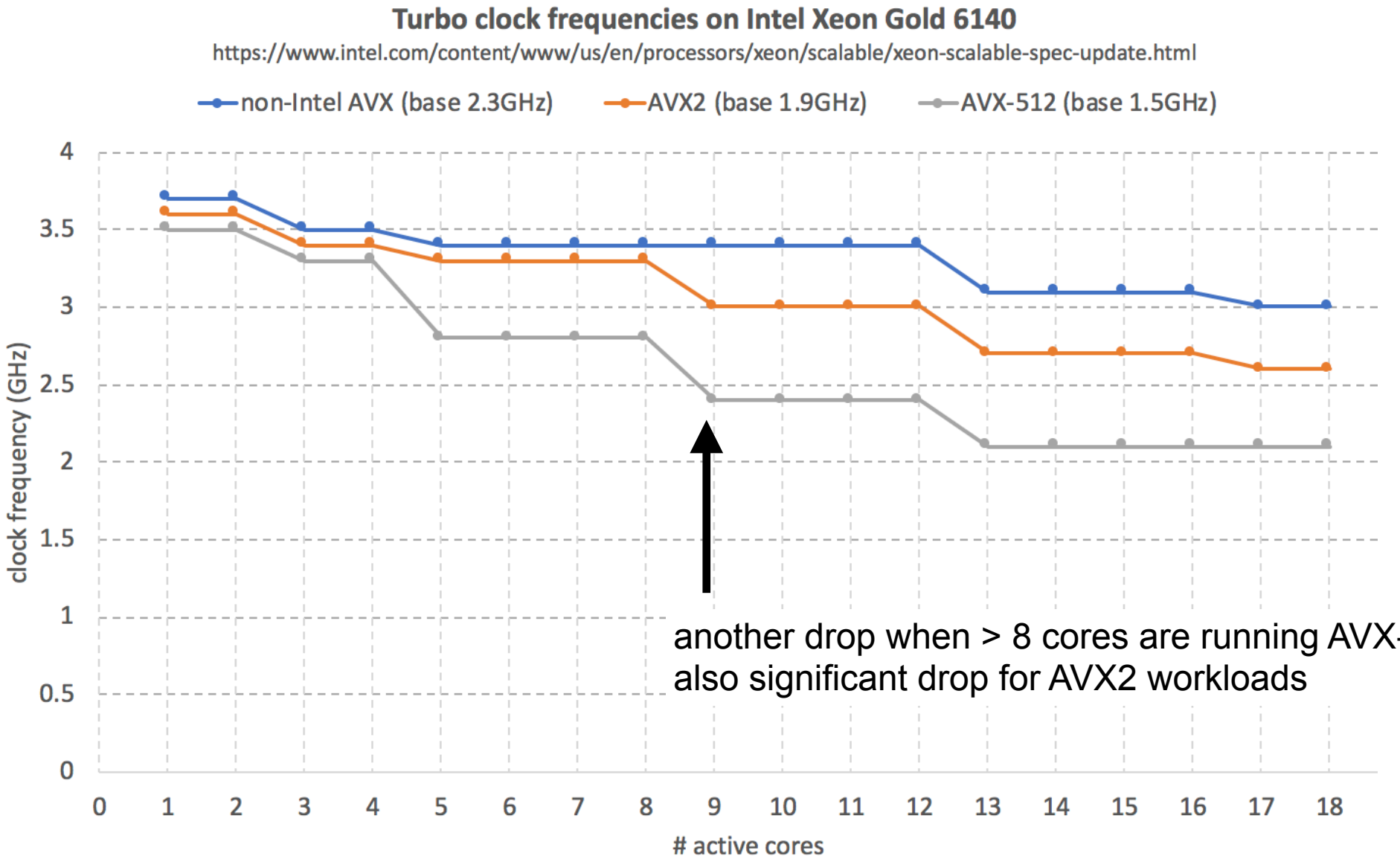
Intel Skylake: AVX-512 vs turbo



significant drop when > 4 cores are running AVX-512 workloads

GHENT UNIVERSITY

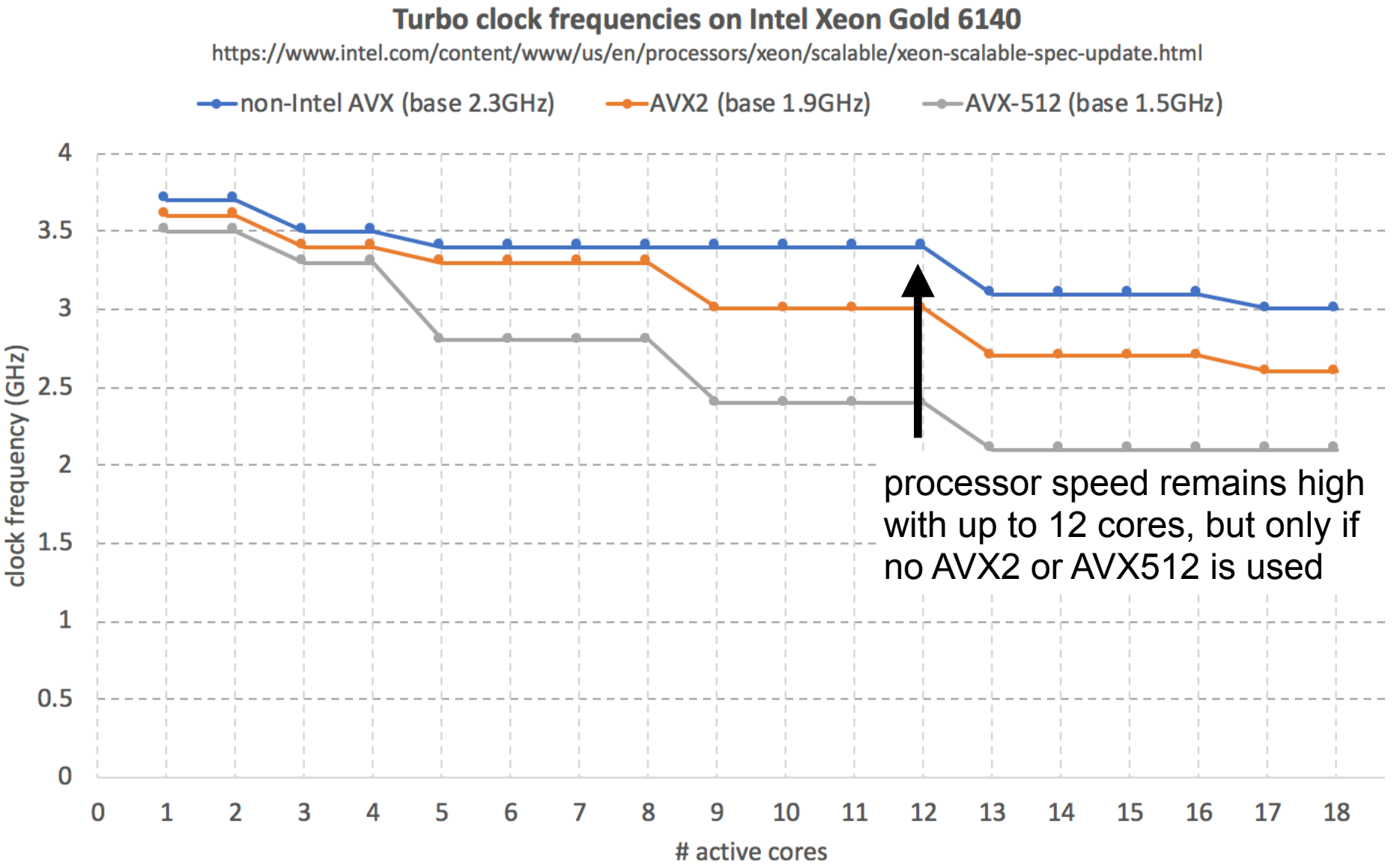
Intel Skylake: AVX-512 vs turbo



another drop when > 8 cores are running AVX-512
also significant drop for AVX2 workloads

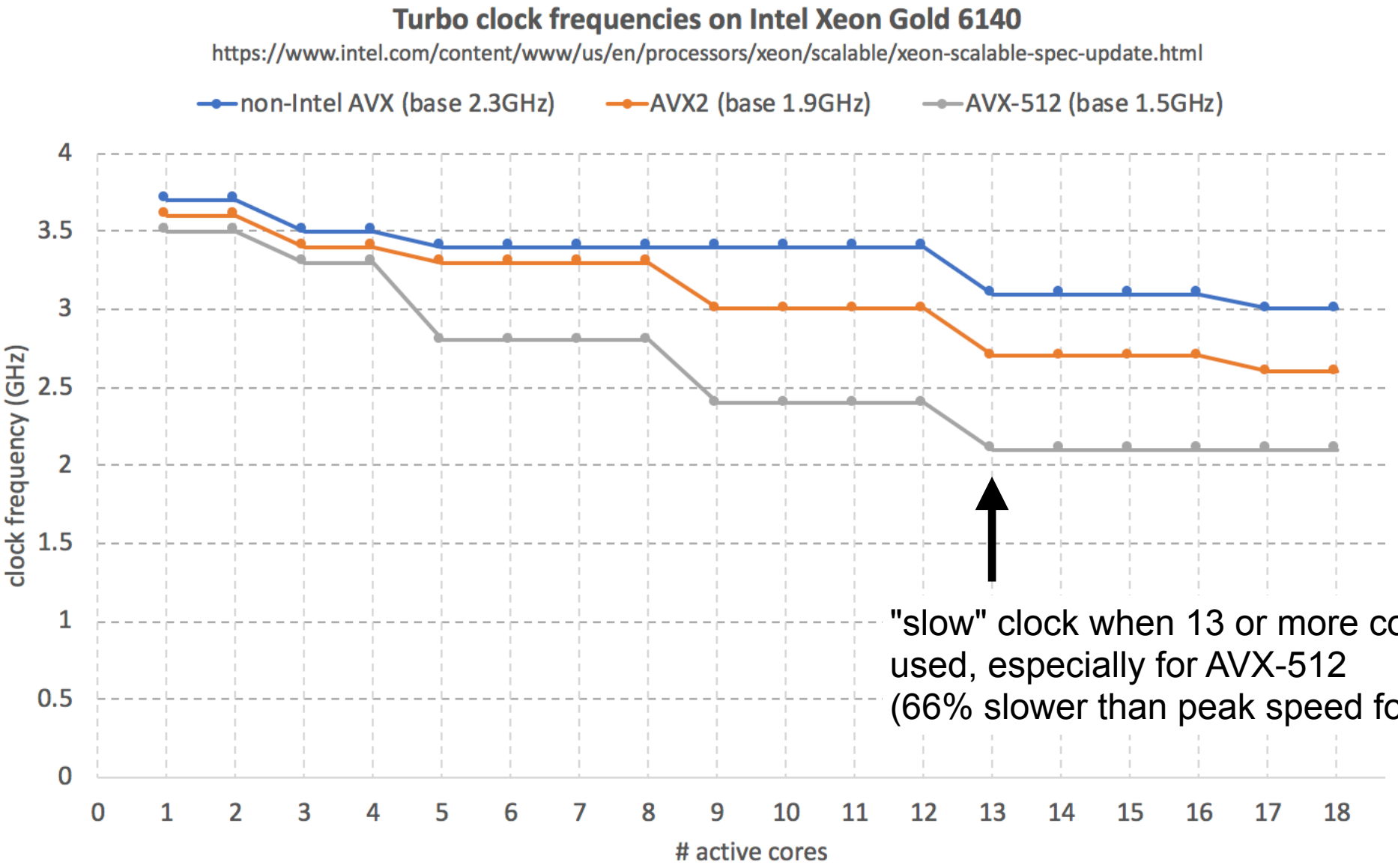
GHENT UNIVERSITY

Intel Skylake: AVX-512 vs turbo



GHENT UNIVERSITY

Intel Skylake: AVX-512 vs turbo



"slow" clock when 13 or more cores are used, especially for AVX-512
(66% slower than peak speed for AVX-512)

GHENT UNIVERSITY

Intel Skylake: AVX-512 vs turbo

- aggressive *per-core* clock speed scaling for (AVX2 &) AVX-512
- unclear if AVX-512 is (always) a good idea
  - more work gets done per instruction when using AVX-512, but ...
  - just a couple AVX-512 instructions can trigger slower clock speed!
- care must be taken when evaluating scaling
  - single-core performance is not a "fair" base when comparing with full nodes
  - not using all cores per node *may* result in better performance...

Is AVX-512 actually worth it?

- We would like you to benchmark whether AVX-512 is beneficial or not...
- Run jobs on skitty/victini using software built for delcatty or golett

```
#!/bin/bash
module swap cluster/golett  # use software built for AVX2
module load CP2K/5.1-intel-2018a
module load vsc-mympirun

mympirun cp2k.opt ...
```

- Report back results to hpc@ugent.be

GHENT UNIVERSITY

Problems or questions?

- contact `hpc@ugent.be`
- make it clear in e-mail subject that it's related to pilot clusters
- provide clear problem description
  - what did you expect to work, what went wrong
  - mention relevant error messages, job IDs, etc.
  - mention location of output files in your account (please don't send them in attachment)
  - exact steps to reproduce the problem