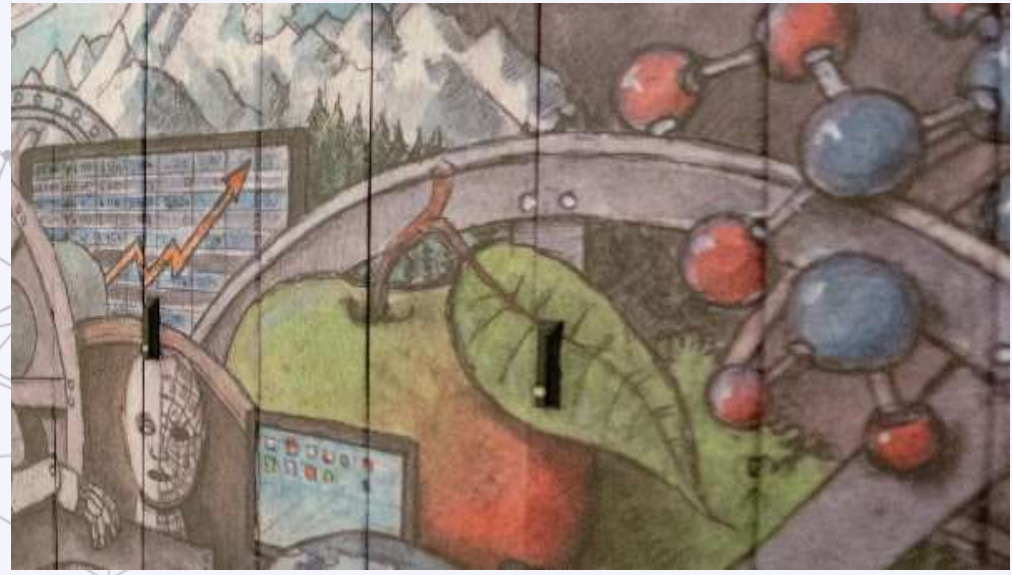
Vlaanderen
is supercomputing

Tier-1 proposals

12/01/2021



VLAAMS
SUPERCOMPUTER
CENTRUM

Innovative Computing
for A Smarter Flanders

vscentrum.be

VSC HPC environment

Later this year

Tier-0
Europe

PRACE

EuroHPC
Joint Undertaking

Vlaanderen
is supercomputing

Tier-1
Flanders

BrENIAC

Tier-2
University

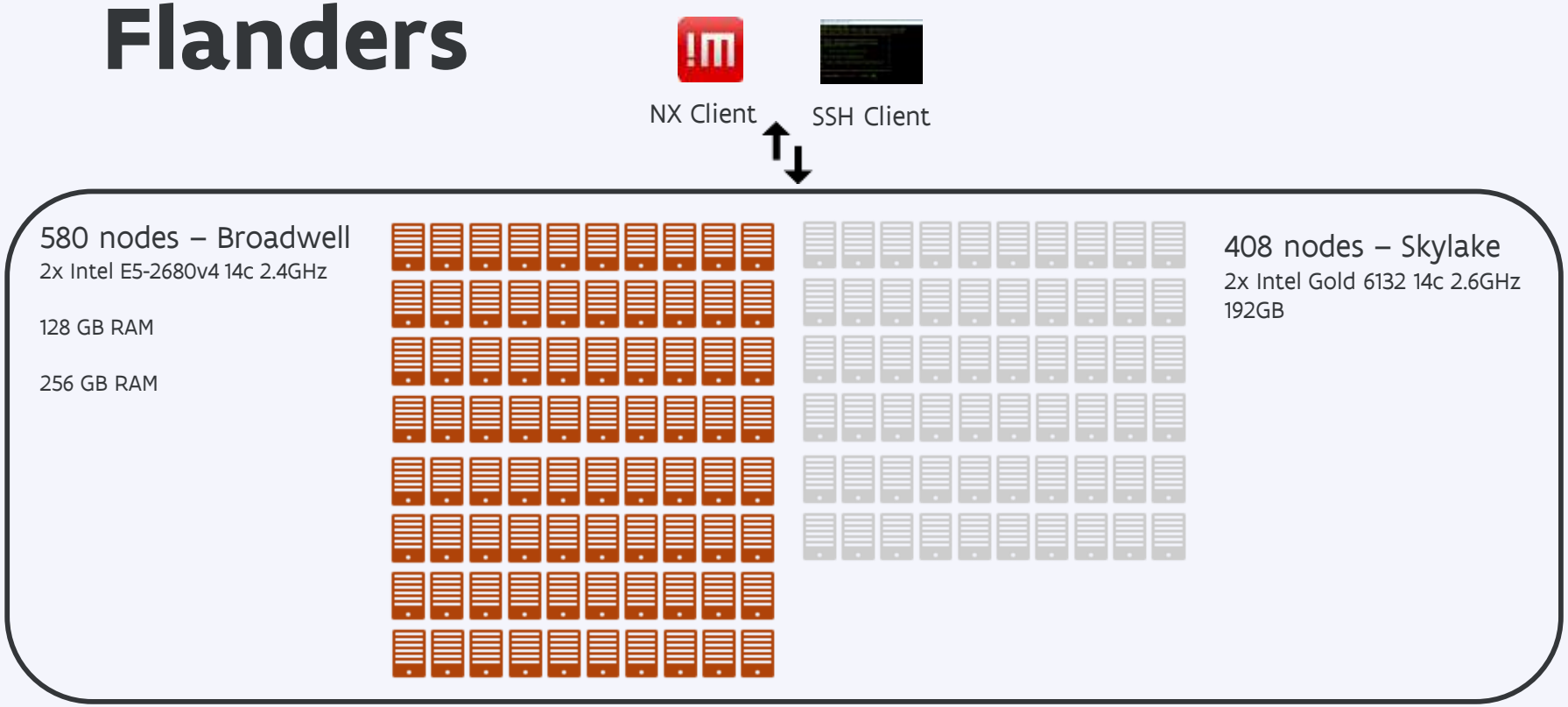
GHENT UNIVERSITY — STEVIN

VUB VRIJE UNIVERSITEIT BRUSSEL — HYDRA

Universiteit Antwerpen — LEIBNIZ/Vaughan

KU LEUVEN UHASSELT — GENIUS

BrENIAC: the fastest supercomputer in Flanders



1.5 TFlop/s peak performance
27664 CPU cores

Check Online VSC doc

Tier-1 Hortense update

Tier-1 node specs

	BrENIAC-Broadwell	BrENIAC-Skylake	Hortense	Hortense GPU
Total nodes	580	408	336	20
Processor type	Broadwell E5-2680v4	Skylake Gold 6132	AMD Rome 7H12	AMD Rome 7H12
Base Clock Speed	2.4 GHz	2.6 GHz	2.6 GHz	2.6 GHz
Cores per node	28	28	128	128
Total cores	16240	11424	43008	
Memory per node (GB)	435x128/145x256	192	294x256 /42x512	
Memory per core (GB)	4.5/9.1	6.8	2 / 4	
Local Disk	128GB SSD	480GB SSD		
Peak performance (TF)	623	950	3300	
GPU per node				4 x nVidia A100-40
Network	Infiniband EDR 2:1	Infiniband EDR 2:1	Infiniband HDR-100	

Tier-1 Hortense – next steps

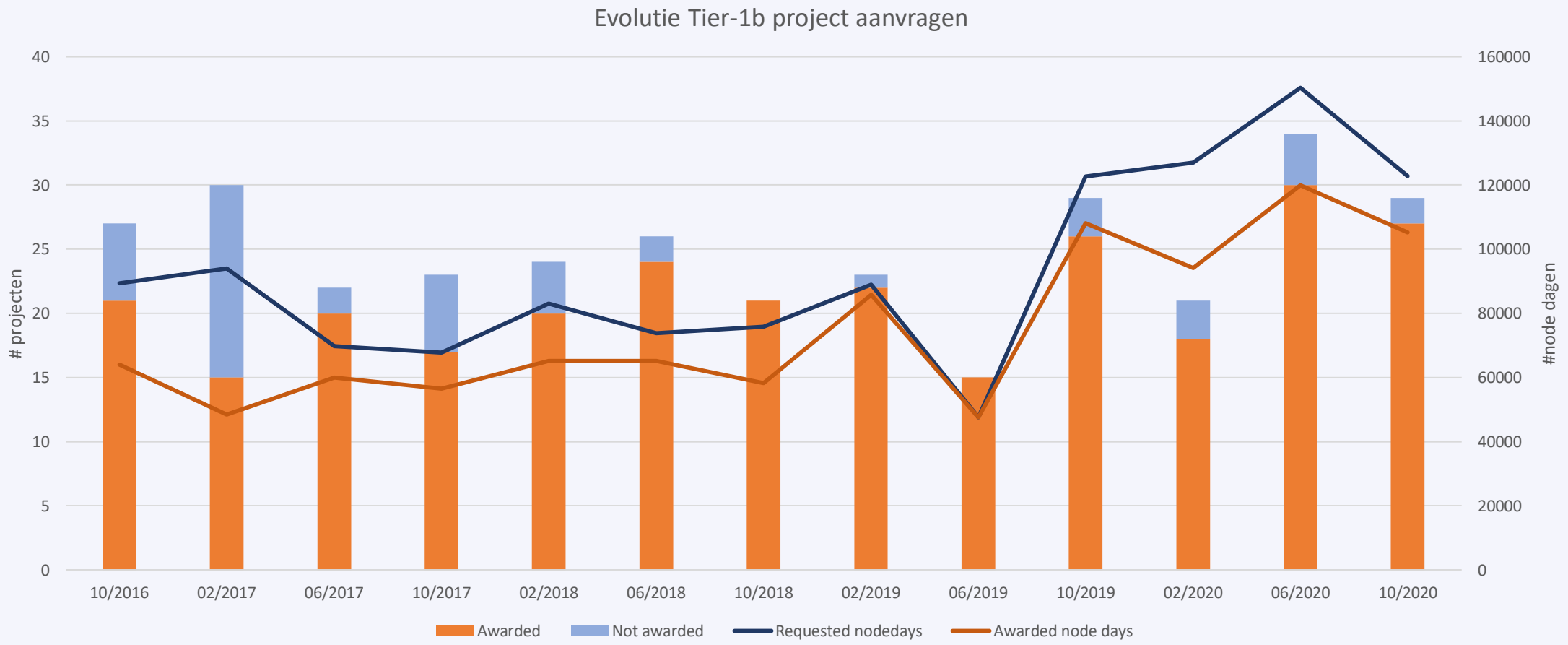
- Installation delayed
- Pilot projects will be possible
- Machine should become operational in next cutoff period
- BrENIAC Skylake will remain operational for 2 more years

LUMI – power beyond our Tier-1

- Large Unified Modern Infrastructure
- Consortium for a pre-exascale machine
- As partner Belgium will get access to the machine
- Lumi kick-off with specifics about the machine on 14/1
- Registration via https://bit.ly/lumi-be-registration.

Back to the proposals

Tier-1 proposals – Evolution



VLAAMS SUPERCOMPUTER CENTRUM

Regulations 2021 general guidelines

Rules and forms on the VSC-website

Use the 2021 template!

9 sections
- Provide all requested information
- Respect the length limits of each section
- Be consistent in the terminology you use

Send us your draft proposal

Feasibility of the proposed tasks
- Clearly describe the tasks
- Support your target job sizes by scaling tests
- Use the resources efficiently
- Be clear on how the tasks are managed

Regulations 2021 – Changes

Job sizes now expressed in core-hours
(no longer in node-days)

Storage estimates now expressed in TiB
(no longer in GiB)

Optional request for GPU hours

Collaborative grants

Februari cut-off is only for BrENIAC
NO GPU requests

Proposal limits

Access	Researchers performing the work | Experienced HPC users
Amount of CPU time	> 500K core-hours | < 5M core-hours, unless justified
Storage	Default: 2TB scratch & 200k files | Describe data management
Duration	8 months | 20% to be used within the first 3 months
System limits	Max 3-day walltime | 128/192/256 GB RAM | Software license

Starting grant limits

Request a starting grant on time!

Access	Individual application
Amount of CPU time	500K core-hours
Storage	1TB scratch & 100k files
Duration	4 months
System limits	Same limits as for proposal

Tier-1 Allocation Board – TAB

Walter Lioen, chairman (SURFsara, The Netherlands)

Nicole Audiffren (Cines, France)

Gavin Pringle (EPCC, UK)

Isabelle Dupays (IDRIS, France),

Robin Richardson (UCL, IRIS, UK)

Every proposal = independently reviewed by 2 board members

During review meeting, rating is discussed with the full board

The proposal should be self-contained

The full board decides on a final rating

Who – Are you eligible?

University or college of higher education
in the Flemish Community

Research institution under the authority
or supervision of a university or a college

University hospital
(Art. 4 law on hospitals and other care facilities,10 July 2008)

SOC: VIB, Flanders Make, Imec and VITO

Higher education,
such as a special university institution
(Art. 169quater of the Flemish Parliament Act of 12 June 1991)

Institution for post-initial education,
support centre for policy-relevant research,
Centre for Research & Development Monitoring

Research institutions of which the activities are entirely or partially financed by the Flemish Community or the Flemish Region

Limited : Federally financed

Who – Are you a Tier-1 candidate?

How – Writing a good proposal

1

Title of research project
Promotor / PI
Needs to proof that it is a scientific validated & financed project

- IWETO/FRIS link is recommended
- Attach confirmation letter
- If project has no scientific approval, letter of the institution

2

Short description: describe the goal of the research

Clear and concise, layman's terms

Indicate collaboration with other groups

The TAB appreciates well-written descriptions!

How – make a good proposal

3

List all users that will use the system

Normally only one proposal per user doing the work

Describe experience of using HPC resources in the past

(Tier-0/Tier-1/Tier-2 infrastructure in Belgium and abroad)

Mention previous granted allocations

4

Describe why the work can not be done on Tier-2

Use checkbox

Or

Specify others

How – make a good proposal

5

Provide information about software that will be used

specify module name when possible

In case an older software version is used mention specific reason

Provide license / Usage agreement in case of not open source software

If a license is needed:

- Prove you have a valid license
- Prove your license allows to install it on Tier-1
- Be sure the license system used can be set on Tier-1
- Be sure you have enough licenses for your biggest job

Use a starting grant!

How – Writing a good proposal

6

Efficiency tests, for each software/task

This is the key to a good proposal

Use a starting grant !

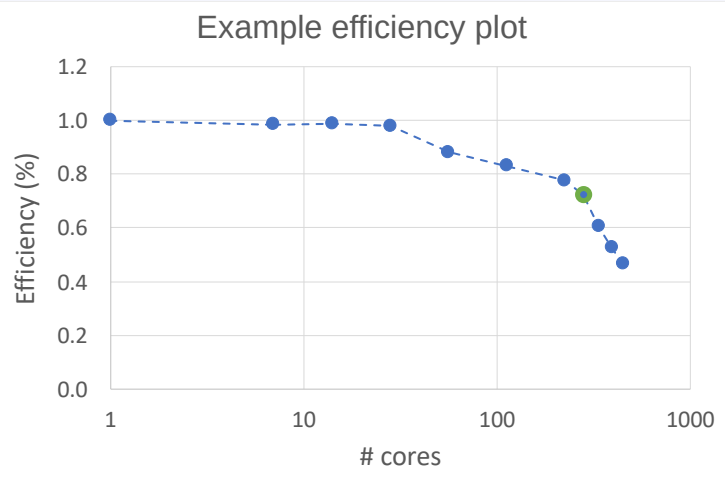
Use the template

Use relevant captions/titles

Use a starting grant!

Number of nodes	Total number of cores	Wall clock time (s)	Speedup (w.r.t. base line)	Efficiency
$A_{baseline}$	$B_{baseline}$	$C_{baseline}$	1.00	1.00
A1	B1	C1	$C_{baseline}/C1$	$(B_{baseline}*C_{baseline})/(B1*C1)$
A2	B2	C2	$C_{baseline}/C2$	$(B_{baseline}*C_{baseline})/(B2*C2)$
Baseline = minimal configuration with which your computational task can be carried out on Tier-1.				
Wall clock time is difference between start/end of the computational task, including any I/O operations as part of that task.				

Number of nodes	Total number of cores	Wall clock time (s)	Speedup (w.r.t. base line)	Efficiency
1	1	4000	1.00	1.00
1	7	580	6.90	0.99
1	14	289	13.84	0.99
1	28	146	27.40	0.98
2	56	81	49.38	0.88
4	112	43	93.02	0.83
8	224	23	173.91	0.78
10	280	19.8	202.02	0.72
12	336	19.7	203.05	0.60
14	392	19.4	206.19	0.53
16	448	19.2	208.33	0.47



VLAAMS SUPERCOMPUTER CENTRUM

Scaling tips – Representative results

- Use a relevant problem size
  - Model size, geometry, atoms in model
  - Grid size, with a relevant workload in each cell
  - Same AMR levels

- Use the minimum time of at least 3 separate runs

Scaling tips – Hardware

- BrENIAC has 2 different CPU types
  - Be sure to use a specific type in your tests
  - Initally test on both types to determine which you can use best
  - Mention the type used in the proposal

- Inter-node communication
  - If timings strongly vary in multi-node jobs, consider using nodes on the same island

    https://vlaams-supercomputing-centrum-vscdocumentation.readthedocs-hosted.com/en/latest/leuven/breniac_quickstart.html#job-communication-and-network-islands

Scaling tips – Parallelization options

- Consider e.g. hybrid MPI – OpenMP parallelization (if possible)

		wall times (s)					
		broadwell	skylake				
# nodes	# cores	MPI	MPI	MPI + OpenMP (2 threads / process)	MPI + OpenMP (4 threads / process)	speedup (wrt baseline)	Efficiency
1	1	48720	**30608**	30608	30608	1	1
1	6	11330	**6042**	6083	6235	5.1	0.84
1	12		**4223**	4494	4664	7.2	0.60
1	24		**2419**	6528	2749	12.7	0.53
2	48		**1494**			20.5	0.43
4	96		**1109**			27.6	0.29
7	192		**961**			31.9	0.17

Scaling tips – What to test

- Start with intra-node scaling
  - Start from smallest core count, preferably single core
  - Use appropriate timesteps/simulation time to fit in 3-day walltime, do not change the domain size
  - Run with 1/7/14/28 cores
- Then move on to multiple nodes
- You can make use of profiling tools
  - Can help you to see how much time is spend on CPU/MPI/IO

Scaling tips – Parallelization schemes

- Also investigate other parallelization options provided by the software

  Example: Vienna Ab initio Simulation package (VASP)

  - k-point parallelization:
    KPAR: number of k-points to be treated in parallel
  - band parallelization:
    NPAR: number of electronic bands to be treated in parallel
  - distribution over plane-waves for cores working on same k-point & band

(Additional tip: "LREAL" flag strongly influences performance and scaling!)

Number of cores	Combinations to test in case of 4 k-points		
1	KPAR=1,NPAR=1		
7	KPAR=1,NPAR=1,7		
14	KPAR=1,NPAR=1,2,7,14	KPAR=2,NPAR=1,7	
28	KPAR=1,NPAR=1,2,4,7,14,28	KPAR=2,NPAR=1,2,7,14	KPAR=4,NPAR=1,7
56	KPAR=1,NPAR=1,2,4,7,14,28	KPAR=2,NPAR=1,2,7,14,28	KPAR=4,NPAR=1,2,7,14

Scaling tips – High-throughput computing (HTC)

- Test scaling in the node
- If multi-core is loosing efficiency, stick to single core
- Test timing with a single work item on a single node
- Test timing with 28 work items on a single node to check if there are bottlenecks (eg I/O or memory)
- Explain workload management, use worker framework

https://vlaams-supercomputing-centrum-vscdocumentation.readthedocs-hosted.com/en/latest/jobs/worker_framework.html

Scaling tips – Presenting the results

- Use the template tables
- Use appropriate titles and captions
- Explain anomalies in plot and table
- Explain chosen job size
- Make sure that you have a data point for the chosen job size

Example table

# nodes	# cores	Wall clock time (s) minimum of 3 runs	Speedup (w.r.t. baseline)	Efficiency
1	1	251561	1	1
1	7	41684	6.0	0.86
1	14	21278	11.8	0.84
1	28	11196	22.5	0.80
2	56	5970	42.1	0.75
3	84	4285	58.7	0.70
4	112	3493	72.0	0.64
8	224	2141	117.5	0.52
10	280	2026	124.2	0.44
12	336	1738	144.7	0.43
14	392	1740	144.6	0.37
16	448	1641	153.3	0.34

Table1: Scalability test for my software on Skylake CPUS, using 32 x 32 x 32 grid cells with 7 levels of refinement (333,824 grid cells in total)

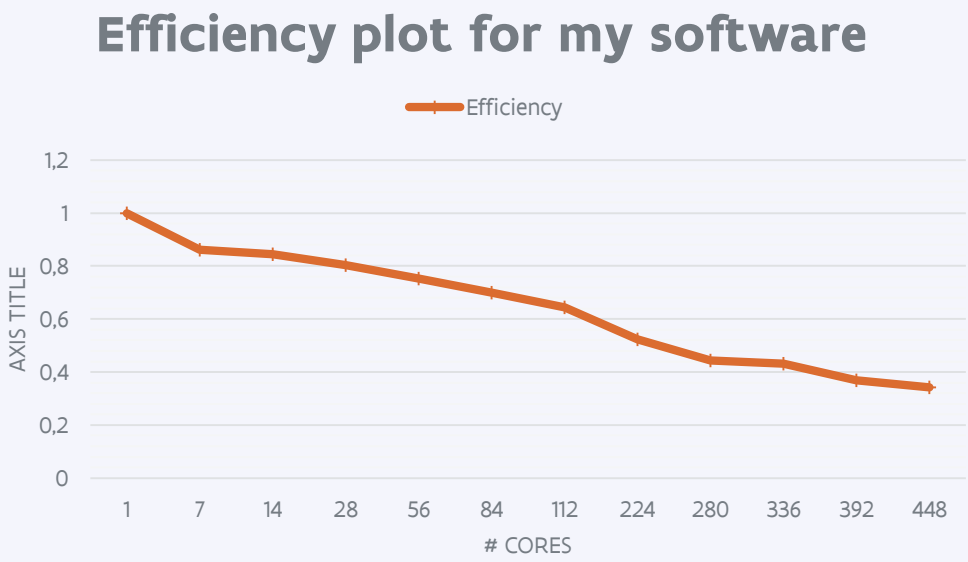
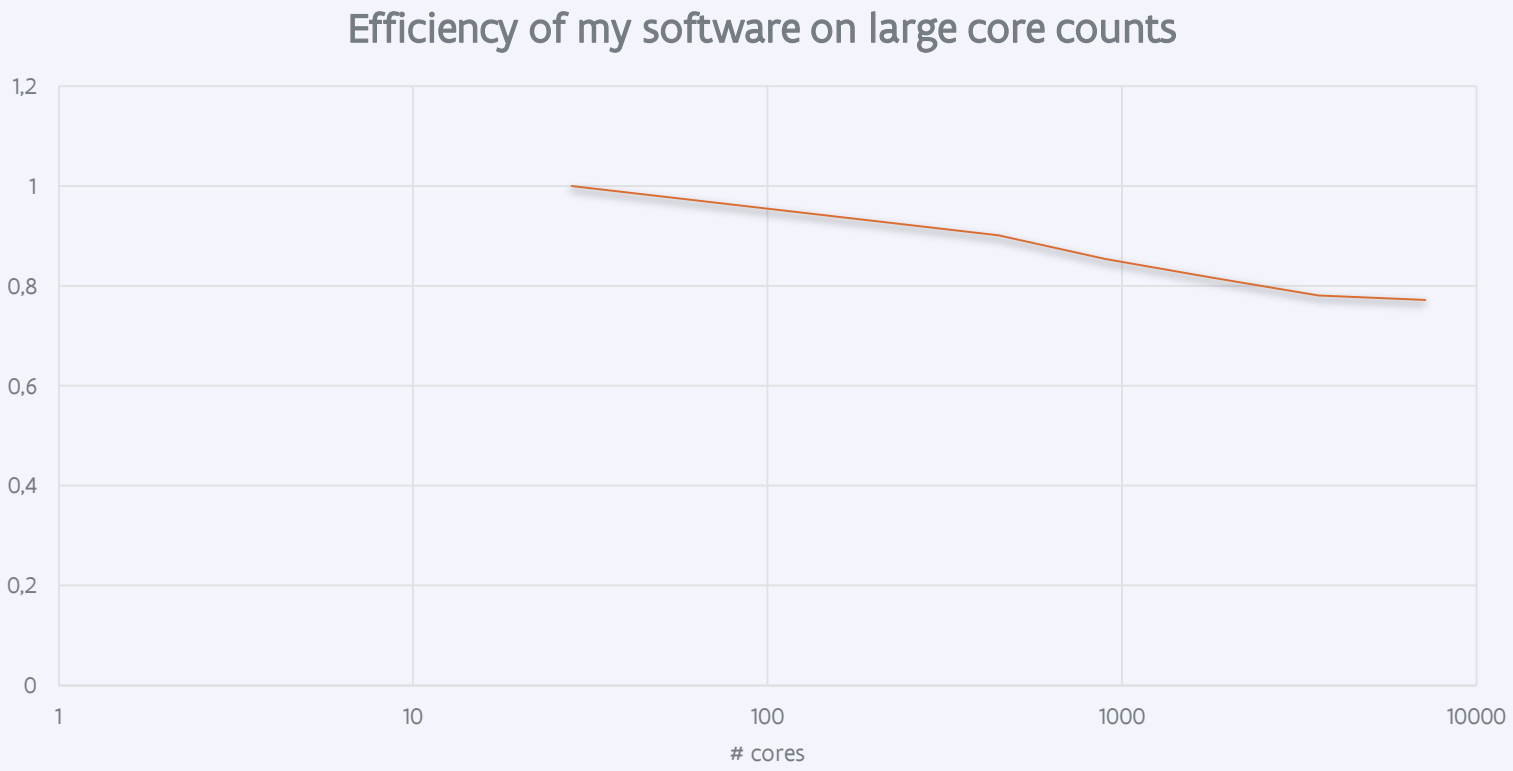


Figure 1: Parallel efficiencies for my software on Skylake CPUS 32 x 32 x 32 grid cells with 7 levels of refinement (333,824 grid cells in total)

Remarks ?

If you would take 6 nodes (168 cores) as your preferred size, also include timing results for 6 nodes

Example table



How – Writing a good proposal

7

Justify the requested number of core-hours, based on scaling tests in (6)
Break down the work as much as possible
Describe well each step and the resources needed, in table and text

Describe scratch space usage if you run tasks concurrently

	Core-hour calculation							Storage volume estimate	
Computational task	Number of such jobs	Wall clock time (in hours) per job	Number of Tier-1 nodes per job	Number of Tier-1 cores per job (per node!)	Total core -hours per task	Memory usage (GiB) per node per job	OpenMP / MPI / OpenMP + MPI (hybrid) / worker framework / atools / etc.	Tier-2 DATA/HOME volume (TiB) + number of files	Tier-1 SCRATCH volume (TiB) + number of files
Task1	A	B	C	D	A x B x C x D				

How – Core-hour calculations

	Core-hour calculation							Storage volume estimate	
Computational task	Number of such jobs	Wall clock time (in hours) per job	Number of Tier-1 nodes per job	Number of Tier-1 cores per job (per node!)	Total core-hours per task	Memory usage (GiB) per node per job	OpenMP / MPI / hybrid / worker framework	Tier-2 DATA/HOME volume (TiB) + number of files	Tier-1 SCRATCH volume (TiB) + number of files
Task A CP2K – molecular dynamics 50 ns runs PBE functional 1 -> 5 water molecules	A	B	C	D	A x B x C x D	Check the math!			
					Sum of core-hours applied for = ..				Largest amount of scratch disk required + number of associated files = ..
Important information									
		3 days is the maximal wall clock time for any job.				Memory limits (GiB/CPU node) BrENIAC: 128->256 Hortense: 256->512			

How – Core-hour calculations

	Core-hour calculation							Storage volume estimate	
Computational task	Number of such jobs	Wall clock time (in hours) per job	Number of Tier-1 nodes per job	Number of Tier-1 cores per job (per node!)	Total core-hours per task	Memory usage (GiB) per node per job	OpenMP / MPI / hybrid / worker framework	Tier-2 DATA/HOME volume (TiB) + number of files	Tier-1 SCRATCH volume (TiB) + number of files
Task A CP2K – molecular dynamics 50 ns runs PBE functional 1 -> 5 water molecules	30	48	10	28	403200	64	MPI	0 TiB 0 files	0.2 TiB 5000 files
Task B CP2K – molecular dynamics 100 ns runs PBE functional 1 -> 5 water molecules	5	72	20	28	201600	64	MPI	0 TiB 0 files	0.1 TiB 500 files
					604800				0.2 TiB 5000 files

How – Core-hour calculations (HTC)

	Core-hour calculation							Storage volume estimate	
Computational task	Number of such jobs	Wall clock time (in hours) per job	Number of Tier-1 nodes per job	Number of Tier-1 cores per job (per node!)	Total core-hours per task	Memory usage (GiB) per node per job	OpenMP / MPI / hybrid / worker framework	Tier-2 DATA/HOME volume (TiB) + number of files	Tier-1 SCRATCH volume (TiB) + number of files
Task example worker MDTraj post-processing 5000 files	10000	0.5	1	1	5000	3 (84 GiB for 28 jobs in one node)	These single-core jobs will be packed within 1 node using worker framework	1 TiB 10000 files	0.1 TiB 5000 files

Assuming you can indeed run 28 tasks on a single node!

Running fewer tasks per node may be necessary in certain cases (e.g. insufficient memory).
Your number of cores per job will then be > 1.

How – Writing a good proposal

8

Describe your workflow
- usage pattern/timing
- task workflow manager
- how you organize in and outflow of data

Don't forget – allocated time should be used in 8 months
20 % used on the first 3 months!

9

If your application is exemplary, can it be used as an example ?

How – Writing a good proposal

Proposals are submitted in EasyChair

You can submit only 1 file

Combine your proposal with attachments in 1 PDF file

How – Using the allocated grant

Be ready to start working as soon as you are granted

Do not wait until the end of the period

Use all the granted time

If applied for a “Starting Grant” -> use it!

It is going to be checked by the TAB
Extensions are not granted (unless for a very good reason)

Collaborative Grants

Collaborative Grant

Rules and form on the VSC-website

At least 3 research groups from different instututes

Application has the same structure as a regular proposal

Allocation at any time
Review by Operation Team

Inform us if you are interested

Proposal limits

Access	Experienced HPC users Collaborations between research groups/institutes
Amount of CPU time	> Max 10M core-hours
Duration	12 months
System limits	Max 3-day walltime | 128/192/256 GB RAM | Software license

Questions