Draft

Report of the “Expert Working Group for assessing the need and outlining a proposal for an extension of the EUROfusion Gateway within the context of E-TASC”[1]

Executive Summary

The Gateway Expert Group has performed a needs assessment to determine the required resources for an extended Gateway.
The Gateway Expert Group recommends:

1. That the size of the Gateway be increased from its present 1344 cores to approximately 4000 cores
2. That the parallel file system be increased from its current 0.6 PB to 2 PB
3. That a virtual infrastructure supplying 32 virtual nodes, each with 4 cores and 32 GB be provided
4. That a modern software development environment be created
5. That appropriate commercial software be purchased and installed, with a contingency fund established for future software requirements
6. That a Long Term Simulation Storage Facility (LTSSF) for storing simulations from across EUROfusion be established

[1] The members of the expert working group are given in appendix A.

Table of Contents

Background

The Expert Working Group (WG) was tasked to:

- Analyse the various requests for small to medium-scale computing platforms that will be formulated by the ACH and TSVV proposals that will be approved for implementation in FP9,
- Analyse the likely need for medium scale computing power in support of WPTE, WPSA, WPW7-X, WPPrIo and WPPWIE, particularly (but not only) of codes developed through E-TASC.
- Analyse the needs in terms of storage of large amounts of data (both from experiments as well as from simulations),
- Identify computing power, storage and architectures required to meet the foreseen needs as well as the necessary level of service,
- Recommend appropriate solutions in terms of computing power, storage and architectures and their possible mix, as well as in terms of location (possibly with respect to their vicinity to a supercomputer) if applicable,
- Analyse possible articulated needs for auxiliary systems (including visualisation) and services, and provide recommendations for technical solutions.

Working Group actions

The Working Group

- Requested and received input from
  - the PIs of the TSVVs and ACHs
  - the experimental work Packages
  - the enabling research projects
  - the Fusion Technology Department
- Held meetings on the 18th and 26th March, and on the 8th, 16th, 22nd and 29th April
- Attempted to clarify what role the Gateway should play
- Prepared a draft of this report and sent it to key stakeholders for comment
- Prepared this report

What roles should the Gateway play?

The working group sees the Gateway providing an enhanced role in the EUROfusion ecosystem, providing

1. A home for joint code development including all of the necessary compilers and libraries
2. A host for EUROfusion scientific software development with repository, trouble ticket and CI/CD infrastructure
3. A home where codes that are developed within the TSVVs, ENRs and other parts of EUROfusion are deployed (as part of their deliverables)
4. A home for EUROfusion developed scientific databases
5. A place where medium scale (non-HPC) codes can be run
6. A place where current simulation results can be stored
7. A home for cloud deployable software (that could be run on the Gateway or in the academic or commercial clouds)
8. A place where cloud software could be run

Recommendations

The working group recommends

1. **For the Gateway**
   a. Expanding the Gateway to have 4000 cores and 2 PB disk space, with a similar architecture as the HPC (if possible)
      i. Including
         1. 4 nodes with at least 1 TB memory (~ 64 cores per node)
         2. 4 nodes with GPUs of the same type as chosen for the HPC (~ 32 cores / node)
         3. 8 front end nodes (~ 64 cores per node)
      ii. Serious consideration should be given to enable bursting to the cloud (academic or commercial), to be negotiated in the future

1. Depending on the nature of the jobs, bursting to the main HPC should also be investigated, and if feasible, implemented

iii. Some of the cycles should be provided as an academic cloud, to be negotiated with partners in the future

1. This would likely be in the context of EUROfusion members moving in the direction of cloud computing

b. That the Gateway be equipped with a high performance parallel file system of at least 2 PB as well as space for the installation of software packages managed by various entities (Gateway Team, ACHs, TSVVs, …)

i. That this disk space be backed up on a daily basis with restorations managed by users, if possible

c. That the Gateway provide the following software packages in addition to those on the current Gateway

i. Matlab and needed toolboxes (including simulink) with a sufficient number of licenses [survey of the community needs will be required closer to the deployment of the new Gateway]

ii. NAG

iii. Mathematica

iv. Totalview or equivalent graphical parallel debugger

v. (A relevant development environment (compilers, profilers and debuggers) tuned to the relevant gateway hardware infrastructure)

vi. The Gateway Team should work with the ACHs to ensure that the necessary publicly available software is installed (via modules or equivalent)

vii. In addition, a contingency budget should be allocated for additional software

d. That the internal network of the Gateway be InfiniBand (or equivalent) with at least 25 GBits/s

e. That the external network connection to the Gateway be at least 20 GBits/second with redundancy

2. For **virtual infrastructure needed for software development, documentation and related services**

a. That, in addition, 32 virtual nodes, each with 4 cores and 32 GB RAM for the provision of the following services (and to allow for future services) be provided

i. Wiki or equivalent documentation services

ii. Issue tracking system

iii. A software repository server (GITLAB, Atlassian or equivalent)

iv. A CI/CD server with ability to run workers on the Gateway and the HPC

v. Database servers

vi. Front-end for querying stored data (from EF databases as well as simulations stored in IDSs)

vii. Front-ends for the FAIR4fusion recommended services

viii. These virtual machines to be available to the ACHs for management, e.g. in with the access to OpenStack dashboard

    ix. It should be possible to run and produce containers (docker and singularity) with a variety of operating systems (and being able to set up e.g. Kubernetes on the provided OpenStack resources)
  b. **That a modern software development environment be installed** on this virtual environment
    i. With a tentative recommendation to centre this around GITLAB (but this should be confirmed after consultation with the ACHs once they have fully started work)
3. **Provide for a Long Term Simulation Storage Facility** (LTSSF) with the following characteristics[2]
   a. Software to support the storage, cataloguing and retrieval of simulations performed on the Gateway, on the HPC and on any other compute facilities used by EUROfusion users
   b. Enough storage for simulations, with options for increasing the space
      i. 1 PB cache, 10 PB tape storage (providing HSM)
   c. Providing long term storage (at least 10 years)
4. Optionally provide nodes with different architectures

If the Gateway is delivered in 2023H2, stakeholders should be contacted in early 2023 or late 2022 to see if any revision of the above recommendations is needed.

Timing

An upgrade of the present Gateway will be needed if the new Gateway is only deployed in 2023H2. This should, at a minimum, include

- An upgrade in the number of virtual machines with management of these machines handled by one or more of the ACHs
- The installation of a modern CI/CD stack including repositories and trouble tickets
- It might be necessary to explore adding a few more batch nodes (compatible with the current nodes)
- The LTSSF is a high priority and this should be implemented as soon as possible.

[2] more details in Appendix D

Additional notes / recommendations

The working group notes

1. That a supplement to the current Gateway might be necessary before the current end of the Gateway (end of 2023)
2. That co-siting the Gateway with the HPC can bring some advantages but is not thought to be necessary
3. That the LTSSF might be a part of the Gateway but might be better separated so that the accumulated simulation data need not be moved every 4 -- 6 years (if the present model of siting the Gateway id pursued in the future)
4. That a governance model similar to the “Gateway User Board” developed under the ITM and continued into WPCD be implemented

Appendix A: Members of the Expert Group

Fusion Experts

- David Coster (MPG, Germany)
- Patrick Tamain (CEA)
- Rui Coelho (IST, Portugal)
- Marco Wischmeier (MGP, Germany, representing WPTE)
- Denis Kalupin (EUROfusion)
- Richard Kamendje (EUROfusion)

Representative of E-TASC Scientific Board

- Pär Strand (VR, Sweden,)

Computer Experts

- Francois Robin (CEA, France)
- Marcin Plociennik (IPPLM/PSNC, Poland)

Appendix B: Summary of Inputs from the TSVVs

TSVV 1		
Computing resources	**Storage**	**Other Specific Needs**
16 cores on a single node 240 jobs per month 3600 Monthly CPUhrs	20 TB long term storage	
Comment: Reduced models and small jobs for testing etc.		

TSVV 2		
Computing resources	**Storage**	**Other Specific Needs**
~10 cores per run, about 100 jobs per month ~30 interactive parallel jobs per month: ~30 x 10cores x 5nodes x 3hours=5000 cpu-hours as well as a few cpu-hours for interactive runs	Intermediate fast storage: 1000 GB Long-term storage: 1000 GB	

Comment: The Gateway has been a precious machine in the past, where mainly ITM/WPCD activity has been done by us, but also occasionally we use it to run CHEASE, which is the natural interface for producing equilibrium inputs for MARS and HYMAGYC, in order to prepare inputs for long, parallel runs on MARCONI. Moreover, all the porting of these software to EUenv (CPOs) and IMAS (IDS) has been done on the Gateway, of course. In this respect, I would strongly suggest that a machine like the Gateway continues to exist, either at CINECA or somewhere in other ACHs, but, anyway, available for all the people who have worked in the past on the EUROfusion Gateway.

TSVV 3

Computing resources	**Storage**	**Other Specific Needs**
Development/ debugging/ testing of edge plasma codes:		
A few nodes are enough in theory as long as they are able to deal with the number of simultaneously connected users, ideally both with CPU and GPU to account for codes running on accelerated hardware	a few 10s of GB of HOME storage per user	needs a full CI /CD environment available on the machine to allow automatic checks of codes to run when commits are made on repositories. Gitlab runners for example (not available on Marconi HPC)
parallel scaling of edge plasma codes :		
These require more nodes as one needs to run up to a large number of cpus to test scalability a few jobs per month, at least 1000cpus, small amount of cpu time (~200kh.cpu / year),	~1TB of storage on a scratch drive for the whole project	hybrid CPU and GPU nodes ideally, as close as possible to hardware of Marconi HPC

workflows to couple edge codes with synthetic diagnostics:		
a few jobs per month, max 100 cpus, small amount of cpu time (~200kh.cpu / year)		IMAS environment setup and maintained
long term data storage:		
	a minimum of ~200TB of secured storage space for the project	
Comment: - an official Gitlab server would also be very useful (instead of having to rely on local partner's infrastructures even for EUROfusion codes)		

TSVV 4		
Computing resources	**Storage**	**Other Specific Needs**
Compare TSVV3 Development/debugging/ testing of edge plasma codes	Compare TSVV3 Development/debugging/ testing of edge plasma codes Potential need for longer term storage	Compare TSVV3 Development/debugging/ testing of edge plasma codes
Comment: For the TSVV4 project medium-sized resources would mostly be useful for development purposes. Since we are doing gyrokinetic or even fully kinetic simulations, most of the work would be beyond of the scope of medium-sized clusters though. Editors note: the working group has dimensioned this requirement as per the corresponding part of TSVV3 for the purpose of estimating the requirements		

TSVV 6

Computing resources	Storage	Other Specific Needs
Compare TSVV3 Development/debugging/ testing of edge plasma codes	Compare TSVV3 Development/debugging/ testing of edge plasma codes Potential need for long term storage	Compare TSVV3 Development/debugging/ testing of edge plasma codes

Comment: For TSVV 6, the requirements are the same than TSVV 3 and TSVV 1 due to the fact

that the codes involved in these TSVV are the same or quite similar ones.

Editors note: the working group has dimensioned this requirement as per the corresponding part of TSVV3 for the purpose of estimating the requirements

TSVV 7

Computing resources	Storage	Other Specific Needs
		IMAS compatibility

Comment: It seems there is no direct interest in a gateway within TSVV7. Some codes require larger systems, for some there are enough resources at the moment. The only inquiry that I have received up to now concerns the possibility to use the Gateway for development and testing of the IMAS compatibility. Since this is mainly the task of ACH4, perhaps this option is already covered by their request.

TSVV 8		
Computing resources	**Storage**	**Other Specific Needs**
	potential option for handling a large-size centralized database (among other options) for JOREK simulations. Such a database would have significant storage requirements increasing over the next five years to the range of several PBs.	
Comment: Medium size computing resources are not needed for JOREK production simulations. In terms of a development platform in particular for integrated modelling, such a system is needed, albeit not many resources from our side. Furthermore, such a system is one potential option for handling a large-size centralized database (among other options) for JOREK simulations. Such a database would have significant storage requirements increasing over the next five years to the range of several PBs.		

TSVV 9		
Computing resources	**Storage**	**Other Specific Needs**

The estimation of resources needed for the ETS are the following: - number of cores: 96 - number of jobs per month: ~ 10 (estimating about 2 days runtime based on current usage) - CPU hours per month: ~ 46080 (estimating about 2 days runtime on current usage)	- Storage: around 100GB, 30GB long term, 70GB fast storage	
Comment: The feedback from the participants of the project (TSVV9 on runaways) is that an extension of the Gateway could be useful for us to keep running the ETS, plus possibly other runaway codes with moderate needs (NORSE, DREAM, ...), although local computers/small clusters have been sufficient so far for the latter.		

TSVV 10

Computing resources	Storage	Other Specific Needs

we will need the EUROfusion Gateway for the reduced scenario modelling in our project. The key numerical tools will be LIGKA/HAGIS and ETS. 100 000 CPU hours per months,	1 TB long-time storage (workspace) 10 TB short-time storage (scratch)	we will need the latest IMAS/py and access to the experiment databases.
Comment:		

TSVV 11		
Computing resources	**Storage**	**Other Specific Needs**

16 cores and 3 days=1152 CPUh. So 2.5 Million CPUh / year. ~208kCpuhr/month 100 000 CPU hours per months,	Need large storage, for IMAS data format: 100 megabytes/run x 10 for IMAS format = 1 gigabytes/run x 2200 runs/year = 2.2 terabytes / year, so over the 5 years of TSVV11 6 terabytes.	Need a system optimized to handle bursty loads. Production run campaigns may be unevenly spread and the system should not have wasted resources when idle, and simultaneously not lead to significant queuing during heavy modelling campaigns. Therefore 30 runs on 16 cores simultaneously have to be possible at all time, so at least 480 cores. (Note that this bursty loads could be handled by the system being embedded within a flexible partition of a larger HPC cluster, or shifting to a cloud deployment strategy). Data storage should Easily/ quickly accessible at any time, including common read access for all users for all integrated modelling simulation output. Login nodes such that 30 users in parallel are assured fast and responsive virtual desktop environments for GUI manipulation and visualization. At the moment, the gateway display/visualization experience is reported not optimum. A boost on such aspects for the login nodes would be welcomed.

Comment: Estimate based on 30 simultaneous users of the most compute-intensive use-cases, e.g. with models like TGLF/QuaLiKiz/NEO/ASCOT/pedestal and SOL-models. 14 inside TSVV11, and others within WPTE/etc. 100 cases each for the 14 TSVV11 users, and 50 cases each for the 16 users outside TSVV11, so 2200 cases / year.

TSVV 12

Computing resources	**Storage**	**Other Specific Needs**
300 jobs per month. The job sizes will vary from serial jobs (single CPU) to several nodes used as test-cases for real HPC simulations. On average 200 000 CPU-hours per month.	The data storage is not expected to be anything extraordinary. A terabyte for the whole project is expected to be sufficient size scale.	- GPU nodes. - A hosted version control system - Container-running capability may be important - In general, to have an easy-to-access joint development platform.
Comment: Please note, that in addition there is need for larger scale simulations with larger simulations, but they are to be performed outside the "Gateway" in a larger system.		

In addition the ACHs have given input to the document where the HPC hubs have not declared any specific needs and where the IM modelling and DATA hubs comments system configuration and data storage/management needs. The latter depends also on whether or not the Gateway will be used as data repository or not. In this case further requreiments capture is needed.

Appendix C: Configuration of the Present Gateway

The Hardware resources of the new Gateway as independent cluster are:

- 9 Virtual Nodes including the following Gateway User Servers:
  - ITM mysql database itmcatalog: itmmysql1.eufus.eu
  - gforge.eufus.eu
  - portal.eufus.eu
  - www.eufus.eu
  - ticket.eufus.eu
- 1 node x3550 M5 with FC HBA as AFS file server of cell eufus.eu
- 1 node x3550 M5 with FC HBA as login node and spare of AFS server
- 4 nodes thinkSystem SD530 as login node with advanced graphic adapter NVIDIA K80
- 16 nodes thinkSystem SD530 as compute nodes with 192 GB of RAM
- 8 nodes thinkSystem SD530 as compute nodes with 384 GB of RAM (2 of which available as debug nodes with dedicated SLURM queue - gwdbg)
- 10 TB of AFS file server storage area
- 15 PB of Backup storage area for AFS/GPFS [for the Gateway and HPC]
- 600 TB of MARCONI GSS as high performance storage area

The job submission system is based on SLURM with the following partitions and quality of service

partition	qos(quality of service)	#cores	max walltime (default)	max memory	priority	Notes
gw	-	1056 (22 nodes)	48 h	unlimited	-	main partition
gw	**(no qos)**(*)	256	48 h	376 GB	0	max 256 cores per user
gw	**qos_gwlong**	144	144 h	376 GB	40	max 144 cores per user , max 2 job per user
gwdbg	-	96 (2 nodes)	30 m	unlimited	-	debug partition
gwdbg	**(no qos)**(*)	96	30 m	376 GB	0	max 96 cores per user

All the above details are further documented in a dedicated Wiki page (*wiki.eufus.eu*), built in addition to a dedicated ticket system web interface (*ticket.eufus.eu*) where all Gateway users can raise trouble tickets of central software issues/requests.

Draft

In preparation for the original Project Implementing Agreement, the list of software flagged as *required* by the user community was the following:

Software	**Mandatory**	**Scientific Libraries**	BLAS, GotoBLAS2, BLACS, FFTW2 + FFTW3, FFTPACK5, GSL, LAPACK, MATHEVAL (depends on guile), MUMPS, PETSC, SCALAPACK, POCO, SLATEC, CERNLIB, PT-SCOTCH, Blitz++, pspline, boost
		Languages	Recent versions of + java (1.8) + ant (>=1.9.6); + python (2.7 and 3.5) (including scipy, numpy, configobj, cython, dateutil, distribute, docutils, h5py, ipython, llvm, llvmmath, matplotlib, mayavi, mdsplus, meta, mpi4py, netcdf4py, nose, numba, petsc4py, pyfftw, pygments, pyqt, pyqtgraph, pyyaml, pyzmq, QTdesigner, sphinx, sumatra, tornado, wxpython); + perl; + ruby; + R
		Recent LINUX	Red Hat, SuSE, Debian, Ubuntu or equivalent + Environment Modules package system + ACL support on the filesystems + pkg-config
		Compilers	Gnu (gcc, g++, gfortran; at least version 5.2), INTEL (at least version 16, at least 5 licenses), G95 (at least version 0.93), cmake
		Static Debuggers	FORCHECK v.14.5.19 or later
		UNIT TESTING SOFTWARE	FRUIT, FUnit

Draft

		MPI	Modern MPI supporting infiniband
		Resources Management System	LSF, Torque/MAUI (free), Torque/MOAB(Basic) or equivalent (SGE or PBS) + UNICORE
		debugger	totalview (preferred - at least 15 licenses), likwid
		profilers	valgrind
		Storage and databases +io	mds+, MySQL, hdf5 + hdf5view, silo, netcdf + "NCO Suite", adios
		Development platforms	Eclipse (with PTP and PyDev plug-ins) Netbeans
		Analysis platforms	MATLAB (at least 20 licences and the following toolboxes: "SIMULINK", "Control_Toolbox", "Real-Time_Workshop" (possibly now known as "Simulink Coder"), "RTW_Embedded_Coder", "Simulink_Control_Design", "Robust_Toolbox", "Identification_Toolbox", "Signal_Toolbox") Scilab Octave
		Remote Desktop	NX (http://nomachine.com/) X2GO VNC
		XML tools	**Oxygen (or equivalent)** XML editor and toolset (version 16 or higher) **JaxFront** (Generate XML Forms and GUIs in Java or HTML) : one licence with support & maintenance package included, owned by EFDA XMLlint XMLstarlet libxml2
		Visualization	VisIt; Vtk; Paraview; gnuplot; xmgrace; mencoder

		Documentation	doxygen vxxx; graphviz
		Software Monitoring	nagios or equivalent Software for monitoring the status of all hardware and key software and services components. This software should allow for easy extension of the tests for new user services
		Misc.	svn git meld screen tmux ndiff (http://www.math.utah.edu/~beebe/software/ndiff/) texlive cloc zeromq
		Editors	Emacs (with nxml-mode, yaml-mode), geany
		Continuous Build	Jenkins (on service nodes)
		Portal and services	JBoss portal, gforge v5.5 and v6.4 (separate server) advanced servers
	Desired	**Scientific Libraries**	MKL libraries and/or ACML libraries
		Analysis platforms	IDL
		Misc	tree, fsr (if using AFS), gsoap, support for docker.io
		Debugger / Profiler	Allinea DDT / MAP Likwid

For the full list of software currently provided on the Gateway see the link https://wiki.eufus.eu/doku.php?id=namespace:software_resources

One should note that, regarding the Python suite, certain desired modules were identified by the Gateway team as “un-installable” which may be detrimental for the overall user experience e.g. on 3D visualization with mayavi.

Appendix D: Long Term Simulation Storage Facility (LTSSF)

Currently there is no easy way for EUROfusion researchers to save simulations for the long term.

JET experimental data going back to 1983 is available, and ASDEX Upgrade from 1991.

Unless the output from codes has been saved in the experimental databases, it is unlikely that simulation results go back that length of time. The Max Planck Society requires that data behind a publication is saved for at least ten years and it is probably safe to assume that other organizations have similar requirements. Individual scientists or groups have implemented partial solutions to this (e.g. the catalogue system at JET or the SOLPS-MDSPLUS database) but there is no solution for the community.

We think the requirements for such a facility include the following:

- That the inputs and outputs of any simulation can be stored
- That the user uploading data be able to set access rights to the data
- That once uploaded, the data should be preserved for at least 10 years
- That the system automatically capture metadata about each uploaded simulation
    - Who uploaded the data
    - When was the data uploaded
    - Where did the data come from
    - How many files were uploaded
    - How many GB of data was uploaded
    - The names of uploaded files
- That the system encourage the addition of other metadata (including provenance data, versions of the code, ...)
- That the system makes the metadata available via an interface allowing searching
- That, subject to permissions, the system provide a way of downloading all or part of the simulation data
- That if IDS's are uploaded (which should be encouraged), the system provides a direct way of accessing data from the IDS's (subject to permissions)
- That there is an effort to ensure that as much of the data as possible is made FAIR, including the issuing of PID's
- One goal would be to be able to reproduce earlier simulations

Some combination of the JET based catalogue system and the SOLPS-MDSPLUS database might be a good starting point for a LTSSF. It might be possible to leverage the SIMDB work that is currently being done for ITER.

Appendix E: Additional input (ENR)

16-64 cores for <6h runs. 100 runs per months (5 cases, 20 time points each) ~50000 CPUhrs per month (for all codes) ~1TB for short term storgae 50 GB for long time storage libraries: NAG, IMAS-python continously synchronised with ITER, connection to experimental databases (preferably via IMAS), MUMPS, PETSc
Computing : - typical number of nodes: 4 - typical number of cores: 192 - number of jobs per month: 20 - monthly cpu-hours: 4 000 - monthly GPU hours: 5 000 Storage : - Intermediate fast storage: 10 TB - Long term needs: 2TB
Computing: typical number of cores: 1-24 number of jobs per month: 100 estimate of monthly cpu-hours needed: 30 000 Storage: Intermediate storage: 200 GB Long term storage: 1 TB
On simulation of reflectometry, main focus of our work in general and of the present EnR project in particular, we use FDTD full wave codes mainly of the REFMUL* family which comprises 2D codes (REFMUL, REFMULX and REFMULF) and a 3D code (REFMUL3). The needs of these codes are quite distinct. The 2D codes, with REFMULF being the present work horse, can definitely benefit from a medium size cluster. 2D REFMULF is threaded using OpenMP and there is a development version using GPU with OpenACC. A typical run using it on one MARCONI node takes from under one hour for the lowest frequency band (K) up to almost 24 h from the highest frequency band (W). Usual projects can take thousand of 2D runs which we fan out, according with resources available and authorised, to several nodes. For the lowest frequency bands it is possible to have several instances of the code running simultaneously on on node. In what concerns the 3D code, its requirements are definitely HPC ones. Requiring from 64-256 nodes, with an average of 128 nodes pre run it needs a HPC like MARCONI. Nevertheless, a medium size cluster can be of help in the setup of a problem, in tests and development of a given new code module and would be welcome. In attachment I include a general description of the 3D REFMUL3 code, for reference (if needed).

I would, at this moment we would point to 15k core-hours/month with a permanent storage of 4TB, with temporary peak storage of 10TB for REFMULF usage. Also, given the fact that we have also a pre-production version of the REFMULF for GPU (OpenACC), 5k gpu-hours would be a plus.
we typically need 240 cores and about 15 jobs in a month. Monthly cpu-hours are 86400. Required storage is about 1 TB.