H2020 EINFRA-5-2015

www.bioexcel.eu

Project Number 675728

D5.1 – Initial business plan

WP5: Governance and Sustainability

Document Information

Deliverable Number	D5.1
Deliverable Name	Initial business plan
Due Date	2016-08-31 (PM10)
Deliverable Lead	FT
Authors	Holger Noack (FT), Ian Harrow (IHC), Alexandre Bonvin (UU), Adam Carter (EPCC), Rossen Apostolov (KTH), Erwin Laure (KTH)
Keywords	Sustainability, user support, services, cost and revenue
WP	WP5
Nature	Report
Dissemination Level	Public
Final Version Date	2016-08-31
Reviewed by	Mark Parsons (EPCC), Rosa Badia (BSC), Berk Hess (KTH), Alexandre Bonvin (UU), Bert de Groot (MPG)
MGT Board Approval	2016-08-31

Document History

Partner	Date	Comments	Version
FT	2016-08-15	First draft	0.1
FT	2016-08-23	Input from IHC, EPCC, KTH	0.2
FT	2016-08-30	Additional input by KTH	0.3

Executive Summary

This document describes scenarios toward sustainability for BioExcel. The described scenarios will be continuously refined during the coming months based on the progress made with our user community. This document should in its current stage thus not be read as a final plan, but rather as a guidance towards development of our final business plan.

The first section presents the vision and mission statements that reflect the shared view of the partners at this early stage in the project. Closely tied into these statements are the success goals presented in the second section. These success goals generate our value proposition and serve as a baseline for our sustainability objectives.

Section three gives a brief overview of the HPC ecosystem. About one-fifth of the European HPC server revenues are related to biosciences, which is the second largest technical segment right after computer-aided engineering. Revenue streams from HPC server revenues in Europe are predicted to grow with 6.9% annually, whereas HPC software revenues are predicted to grow with 7.2% annually until 2018.

Section four presents BioExcel's value proposition to different stakeholders in the biomolecular and HPC ecosystem. The key value that we provide is the enhancement of efficiency, scalability and usability of open-source software important to the biomolecular community. In addition, we will provide free support, including free training to non-profit users. These support offerings are accompanied by premium service offerings to paying users. Annual prices for access to a premium web portal with help desk are suggested to be €120 for non-profit users and €700 for industrial users. In addition, the intention is to provide different consultancy services, which are presumed to be most relevant to industrial users.

Section five then presents a suitable governance structure that is driven by user needs, and gives a first hint on what needs to be considered when choosing a legal vessel.

In section six, three different scale cost scenarios are compared. The common aspect of the scenarios is that revenue streams from our premium service offering allow reduction of our funding rate by partially covering costs of center operation and free user support.

Section seven concludes this business plan with sustainability objectives that support our development towards achieving sustainability.

Contents

1 Vision and mission

Vision: *"A central hub for biomolecular research software"*

BioExcel aspires to become a central hub of biomolecular computational excellence, where users, software developers, computational experts, and resource providers are linked together. Through BioExcel, exchange of knowledge and collaboration between these different stakeholders is facilitated.

Mission: *"Improving efficiency, reliability and ease of use of biomolecular software, and sharing best practices and expertise with software users and stakeholders."*

In doing so, BioExcel will promote wider usage of computer simulations among the biomolecular research community, increase competence of users and quality of computational research performed.

The initial vision of the CoE will focus on biomolecular simulation, and workflows which support these simulations. This is reflected in the current project's pilot codes. However, the longer term vision of the CoE is expected to include computational research for processing and analysis of relevant data. Whereas in the past, pre- and post-processing and analysis have often been less computationally challenging, it is now more common for these parts of a workflow to be performed in parallel and at scale. There is a growing overlap between the techniques that are required to incorporate simulations into wider workflows and those that are used in other computational workflows such as high-throughput sequence analysis, currently considered in use case 1 of work package 3. There are interesting problems to be addressed here such as varying-level parallelism and integration of workflow systems with HPC resources and data services.

2 Success goals

The current BioExcel partners have identified 14 long-term success goals, which according to their emphasis can be grouped into six different categories related to users, software, support, services, partners, and organisation. The first four of these categories generate our value proposition and have to align with stakeholder needs.

USERS

- **A1: Diverse user community**: BioExcel should attract users from e.g. different biomolecular disciplines, including both users experienced in computational techniques and entry-level users.

- **A2: Active user community:** A large number of users should actively contribute to BioExcel to help create attraction and visibility.

SOFTWARE

- **B1: Well-documented, state-of-the-art flagship software**: The open access flagship software provided by the BioExcel partners should be highly usable and up-to-date with the possibilities provided by modern e-infrastructures, particularly with efficiency and scalability.

- **B2: Provision of interfaces, workflows and platforms:** These tools should complement the flagship software to increase user-friendliness and support analysis of performed simulations.

- **B3: Mechanism for inclusion of user feedback:** Software developers should receive direct user feedback to be built into the design cycles for their software. Provides ability to increase functionality and usability of software, while promoting it to a wide community.

SUPPORT

- **C1: Expertise sharing:** A platform for knowledge exchange among experts and users that incorporates and helps to drive basic needs as well as top-level science adequately. Important is competence matching, i.e. the ability to provide relevant answers to different types of questions.

- **C2: Mechanism for easy collaboration:** BioExcel should provide means to simplify collaboration among BioExcel users as well as between users and BioExcel partners.

SERVICES

- **D1: User training:** Provision of use-case driven training, both face-to-face and online, enabling users to make the most of software and workflows. Training offerings need to match different user competencies, and ideally should cover all the possible needs of researchers engaged with computations.

- **D2: Enterprise quality services:** Ability to meet industrial requirements when needed. This regards consulting on modelling, practical support, specialized training, and customized installations.

PARTNERS

- **E1: Work satisfaction:** Experts who to a large extend are involved with provision of support and consultancy should be recognized for the excellence of their work and be provided with a career path.

- **E2: Advanced training:** Continuously advance the competence of the BioExcel experts, which will also benefit the partner organisations.

- **E3: Career opportunities:** Support access to new research and development projects for individual partners.

ORGANISATION

- **F1: Strong, recognized brand:** A professional online presence with responsive helpdesk and active marketing will improve visibility of our software, services, and partner institutions.

- **F2: Stable economic situation:** Ideally, it should be possible to continuously finance user support, basic service offerings, parts of the software development, as well as legal and administrative costs through a mix of grants, membership fees, sponsorships, and commercial fees.

3 HPC ecosystem

Simulations of biomolecular systems are particularly computationally intensive and require parallel computing techniques to achieve timely results. Depending on the mathematics of the underlying science these computations either need to be performed as tightly coupled parallel jobs on high-performance computing infrastructures (HPC), or can be performed as loosely-coupled tasks on distributed networks of processors known as high-throughput computing (HTC).

The two main actors that regularly perform global market studies on HPC technologies are Intersect360 Research (IS360R) and International Data Corporation (IDC). IDC has twice been commissioned by the EC's Directorate-General for Communication Networks, Content & Technology to conduct HPC market studies. The first study was performed in 2010 [IDC_HPC_2010], and a second one during 2014-2015 [IDC_HPC_2015]. The latter study is the most recent study available on the European HPC market, it contains updated market figures up to 2013, and predictions for the years 2014-2018. This IDC study was used as the main source for market data presented in this section.

3.1 HPC hardware in Europe

The worldwide largest buyer of HPC systems are the U.S., which historically purchased just under half of all sold HPC systems. In 2013, Europe purchased circa 27% of the globally sold HPC systems (see appendix). The average annual increase of server revenues in Europe for the years 2009-2013 was 5.7%. For the period of 2013-2018 Europe's compound annual growth rate (CAGR) is projected to be 6.9%, meaning that in this case Europe would be the fastest growing market globally.

3.2 HPC software in Europe

HPC software generated global revenues of €4.4 billion in 2013 [IDC_HPC_2015]. Assuming that the European portion of spending in this market follows Europe's spending in the global server market, gives an estimated €1.2 billion revenues for HPC software. It is estimated that the European spending in this market will expand to €1.7 billion by 2018 [IDC_HPC_2015], which translates into a CAGR of 7.2% for the period of 2013-2018. Apparently, freely available open-source software is not accounted for when looking at HPC software revenue streams. It is thus unclear if usage of open-source software follows a similar or an even accelerated growth.

In a 2014 survey with 97 HPC users, 68% of foresee a need for a 10x increase in performance over the next five years, and 57% even wish for a 100x improvement of performance [Solve_2014]. Lack of support to improve scalability of software is seen as the number one barrier to achieve this goal. This indicates that open-source software is widely used in the HPC community, since a lack of support and maintenance is the biggest issue for open source software.

3.3 HPC and data analytics

The increasing use of HPC infrastructure for high-performance data analytics (HPDA) is driven by the need to tackle data-analytic questions that are complex or time critical [IDC_HPC_2015, p50]. HPC technology allows to tackle more complex questions on data infrastructures by enabling the simultaneous processing of large amounts of data with many variables, thereby effectively shortening time-to-value. Since HPC clusters are designed as multi-purpose platforms, they also allow to pack the entire analytics pipeline (e.g. stream processing, data mining, interactive querying) into a single, cost-effective cluster. Notable HPDA initiatives within Europe include:

- EUDAT offers common data services through a geographically distributed network of general purpose data centers and community-specific data repositories. These shared services and storage resources are distributed across 15 European nations.

- The Research Data Alliance (RDA) is a community focused on "building social, organizational and technical infrastructures to reduce barriers to data sharing and exchange". EUDAT drives a number of RDA working groups, e.g. the Data Foundation and Terminology WG or the PID Information Types WG.

- Big Data Europe designs and evaluates Big Data aggregator platform infrastructures to take advantage of latest European RTD developments.

The worldwide revenue for HPDA servers will grow with a CAGR of 23.5% during the period 2013–2018 [IDC_HPC_2015, p51]. This is more than three times the forecasted growth rate of the worldwide HPC server market.

3.4 HPC technology trends

The HPC technology trends presented below are chosen for their potential importance when formulating objectives for this sustainability plan. BioExcel's technical work packages that deal with software development, libraries, and middleware are following technical trends to greater detail.

Coprocessors/Accelerators: During 2013, X86-based processors from Intel and AMD accounted for 90.3% of the 9.6 million processors installed in HPC systems. The small but growing market for accelerators represented 3.4% of all HPC

processor parts shipped in 2013 [IDC_HPC_2015]. Accelerators should stand for more than half of all new HPC systems in 2015 [IS360R_2015].

Cloud Computing: The proportion of sites employing cloud computing has steadily grown from 13.8% in 2011, to 23.5% in 2013, to 34.1% in 2015 [IDC_HPC_2015]. Here the challenge is about the ability of clouds to handle more types of HPC jobs over time, which contrasts with cloud data storage services where security of data and its geographic location is the biggest barrier to adoption. Some of these challenges, such as GPGPU computations in a cloud environment, are being addressed in the context of the INDIGO-DataCloud project, to which BioExcel is connected through one of its partners.

3.5 Biosciences in HPC

In 2013, the four largest buyer segments that purchased HPC servers were government (27%), academia (20%), computer-aided engineering (12%) and biosciences (11%). Biosciences is thus the second largest 'technical' segmentation in the HPC server use.

Per definition used in the market study (appendix), the bioscience segment only accounts for servers exclusively purchased to run bioscience applications. However, servers bought by users from the government and the academic segments are in their majority multipurpose, and certainly also used to run applications in bioscience or other areas such as computer-aided engineering.

To estimate the overall percentage of HPC server revenues related to biosciences, we assume that the HPC server revenues from the two economic segments government and academia can be further divided into application segments that mirror the distribution of application segments found outside of these two economic segments. This assumption allows us to 'remove' government and academia from the total of server spending (i.e. 100-27-20), and then adjust the percentage of server spending for bioscience to the new total (i.e. 11/0.53). In this assumption, the overall amount of server revenues related to biosciences is 21% throughout all economic segments.

From here it is a small step to assume that the percentage of HPC server capacity used for bioscience software follows the portion of HPC server revenues related to biosciences, i.e. about one-fifth of the European HPC infrastructure capacities is used by bioscience applications.

Within industry, the bioscience software is sourced as 43% from commercial vendors (ISVs), 37% is open-source, and 20% is in-house [Solve_2014].

3.6 Discussion

In 3.1 we have stated trends in server revenues that reflect the spending of buyers. However, in principle the stated numbers could be due to increasing prices of server components, as opposed to increasing volumes of servers bought. Better descriptors to capture the growth of HPC infrastructure, and

compare different global regions, might be the number of cores installed or the maximal LINPACK performance achieved. It is possible to find that information for the top 500 server installations worldwide at top500.org, and more detailed information will be compiled if deemed necessary for this business plan.

The growth of software revenues reported in 3.2 indicates that software is expected to grow at a larger rate than hardware. However, this includes all types of parallel software, and the growth of software use within the biomolecular research community might deviate considerably from the above stated number.

In general, sections 3.1, 3.2 and 3.5 together provide a first qualitative picture, showing that biosciences indeed is a major player in HPC, and how the entire industry might develop over the next two years. The forecasted growth of HPDA reported in 3.3 shows that BioExcel needs to investigate the importance of this technology for the biomolecular community to understand the need for HPC in data analysis.

4 Value proposition

In general, BioExcel will provide value to different stakeholders such as software application users, independent software vendors, or HPC infrastructure providers. The value proposition is generated by our initially presented success goals and has its impact focus on biomolecular researchers that use open-source software. In our value proposition we distinguish between freely available support, and premium services.

4.1 Users

Users from the biomolecular research community can be segmented based on their environment, i.e. government, academia or industry. In some sense orthogonal to that it is possible to distinguish users by biomolecular research area, which makes sense if those differ in their use of parallel computing.

Government users: In government funded organisations, computational techniques are used by production users and research users. Production users tend to have time sensitivity in their operations and hence require stable systems that are easy to maintain and scale. Provision of proven workflows and software interfaces would be a benefit to these users. Research users, in contrast, are able to trade-off system reliability for higher levels of performance, and are often familiar with open-source applications. These users have similar needs as academic users described below.

Academic users: Academia is the major player in open-source software efforts [IS360R_2011] and has in general internal support from technology-affine graduate students, post-docs and in some cases from computer scientist. Academic researchers engaged with computational techniques are usually experienced in how to use their tools, and are well connected to find support if needed. These user types would benefit from improved software scalability and efficiency, which would allow them to tackle increasingly complex questions and

decrease time to results. A different type of user would be entry-level researchers that do not regularly use computational methods. Here, BioExcel could provide value through provision of case studies, guidance and basic training.

Industry users: Shortening time to market is generally the main driver for industry, which is also true when it comes to parallel computing technologies [Solve_2014]. Development also has to show acceptable costs for production of the treatment at scale for the market. Industry users working with development require tracking control of software to validate production of a data package to demonstrate efficacy and safety, and hence tend to use commercial software solutions that allow their research to be performed accordingly. Here, BioExcel could provide value by promoting open interfaces and programming policies, which in turn will facilitate co-development of commercial software by ISVs. Open-source software is more likely to be used by industry for research purposes, where none of the regulatory restrictions of development apply. Another way to provide value to industrial users is to offer customized application support for code conversion and optimization [IS360_2011].

4.2 Other stakeholders

Apart of software users, other stakeholders to which BioExcel will provide value are e-infrastructure and software providers.

e-Infrastructure providers: HPC/HTC infrastructure providers operate the hardware, install and maintain the software (system administrators), and also offer training and helpdesk services to researchers [EIRG_2012]. The quality of the software is part of the user satisfaction concerning the services provided, and also impacts the efficiency on how the infrastructure is used. The value that BioExcel can offer to e-infrastructure providers is to ensure that there is a match-up between the available computing infrastructures and the development of parallel software. In addition, BioExcel could assist system administrators in deployment and testing of software and updates.

Independent software vendors: ISVs for high-end computing applications tend to be specialized, although there is some overlap in broad categories such as molecular modelling. This specialization results in a reduction of competitive pressure that slows down adoption of software to meet customer requirements for greater scalability [Solve_2014]. The majority of ISVs are small or medium-sized enterprises that also must deal with bug fixes, development of new features, and support for a combination of operating systems, distributions, processors, and middleware environments. By implementing a policy of open and non-restrictive programming interfaces and support libraries, BioExcel would ensure that software created within the open-source community allows for ISVs to provide added value, i.e. to develop complementary commercial variants and related software tools. Partnerships with BioExcel could be an interesting commercial opportunity for “co-design” of academic community and industrial applications software. This would require a collaborative environment that encompasses definition and adaptation of programming models to be used

with scientific software, as well as the computing environments and middleware within which those applications will be deployed.

4.3 User support

Our support offerings will be freely available to the entire user community, and are due to their scope expected to be funded by public funding.

Software development: As mentioned in section 3, the lack of support and maintenance is the biggest issue for open source software. The goal of the software development is to improve the efficiency, scalability and ease of use of BioExcel's flagship software, and to provide workflows and interfaces to match the possibilities provided by modern e-infrastructures. This support offering also includes provision and improvement of current web portals like the HADDOCK one by UU and the MDweb from IRB. Finally, documentation needed to support co-development through the community or ISVs will be made available on BioExcel's website. This support offering aligns with success goals B1 and B2.

Software deployment: BioExcel will assist public e-infrastructure providers throughout Europe with installation of its flagship software to guarantee a smooth user experience. This support offering aligns with success goal B2.

Training: Training offerings provided as a public support will be offered only to non-profit users, i.e. user from government and academia. The current idea is to provide workshops or seminars on general topics aimed at specific levels of user expertise. Details of the various training offerings are carved out in work package four. This support offering aligns with success goal D1.

Website: The website is the central part of BioExcel's public face and needs permanent work to stay up to date and remain relevant. All support functions described below are made available through the website. To stimulate frequent visits, the website will provide research & technology news that overlap with interests of the biomolecular user community. Users will also be informed about upcoming conferences, new tutorials and webinars. One section should deal with transparency, describing to users our priorities, how decision processes are made, who our collaboration partners are, as well as presenting board members, advisors, and key personnel. Finally, the website will host a technical section with documentation of our software and recommendations regarding interfaces, libraries, middleware, and best-practices relevant for co-development. This support offering aligns with success goals E3 and F1.

Feedback portal: The website will provide an interface where users can report bugs, problems with performance, and also leave suggestions for desired improvements in scalability or functionality for our core software applications. The feedback will be channelled to BioExcel's software developers. This support offering aligns with success goal B3.

Webinars: Webinars are published on a regular basis, preferably each fortnight. A roadmap of upcoming seminars will be public, whereas past webinars are made available in a searchable manner. To reduce the burden on BioExcel,

speakers from the user community will be invited to present some webinars, i.e. in this case BioExcel will provide the webinar tool and the publication platform. This support offering aligns with success goals A2 and D1.

Tutorials: Tutorials will be in the form of use cases linked together with short video clips to provide "how-to" descriptions to different user segments, e.g. directed at entry-level users, or on a specific topic for experienced users. Active members of the user community are invited to contribute. This support offering aligns with success goal A2.

Matchmaking: To facilitate collaboration, we will build and maintain a simple database with records of the expertise and research interests of active users, PIs and key opinion leaders (KOLs) from both biomolecular and the computational research fields. All too often opportunities are lost due to a lack of knowledge of existing contacts or a missing introduction. This database could be kept internally, only to be used by us to assist users that are looking for collaboration partners with particular competencies and interest. Alternatively, it could be provided as a web service similar to brightowl.pro. This support offering will be very useful for consortia building in preparation of grant applications, and aligns with success goals C2 and F2.

Discussion forum: BioExcel will host a discussion forum (currently ask-bioexcel.eu) where users and experts can help each other and exchange ideas. Different sections for specific topics or different level of experience are made available. To ensure quality of the discussions, each section in this forum needs to be overseen by an BioExcel expert, who if needed can provide input. This support offering aligns with success goal C1.

4.4 Services

Services are a premium offering to paying users. They are expected to hold enterprise quality and require contractual service level agreements.

Training (premium): Experience from the free training offering provided to non-profit users will allow us to develop premium offerings tailored to the specific needs of paying user segments. It is likely that these premium offerings will be held in smaller groups to allow for a larger extend of supervision and hands-on exercises.

Web portal (premium): The premium web portal differs from the public web portals by the possibilities of customization, and provision of SAAS offerings. If possible and desired by the users, e-infrastructure providers will be connected to this platform so that users do not need to deal with allocation of computational resources. This support offering would need to be developed, and if implemented would align with success goal D2.

Help desk: The helpdesk must be manned during core business hours to allow direct response to user requests. The expected requests could be related to setting up submission files, advice on methodology, and troubleshooting. It might

be reasonable to integrate this help desk into the premium web portal. This support offering aligns with success goals D1 and D2.

Consulting: Consultancy could include customization of software, development of particular software functionalities, or co-development with ISVs. On top pf that consultancy assignments could include performance of contract research. Commercial clients are likely to pay for this service with their own funds. Another type of customer can be public organisations that have allocated funding resources to acquire such consultancy services on a commercial contract basis. This support offering aligns with success goal D2.

4.5 Service packaging

We aim to offer our services through subscription fees and service fees. Subscription fees are based on a membership and are preferred from our perspective, since they provide greater stability by allowing to predict future revenues and corresponding workloads. A membership will thus have additional benefits as compared to service-fees. This benefit can be in the form of a bundled service offering at a lower price point as compared to using the same services through service-fees. Hence, we propose the following packaging of our services:

1. Annual membership fees, which include a fixed number of premium training days, plus permanent access to the premium web-portal and help desk.

2. Service fees, which require separate payments for premium training, access to the premium web-portal, or consultancy.

5 Governance and legal structure

Enhancing and maintaining scientific software, and providing free support and premium services, requires a sustainable governance structure with clear interfaces towards users and other stakeholders. This governance structure must be embedded in an organisational form that provides flexibility to interact with different types of users on a daily basis, while also accounting for existing restrictions in external engagement from the BioExcel partner organisations.

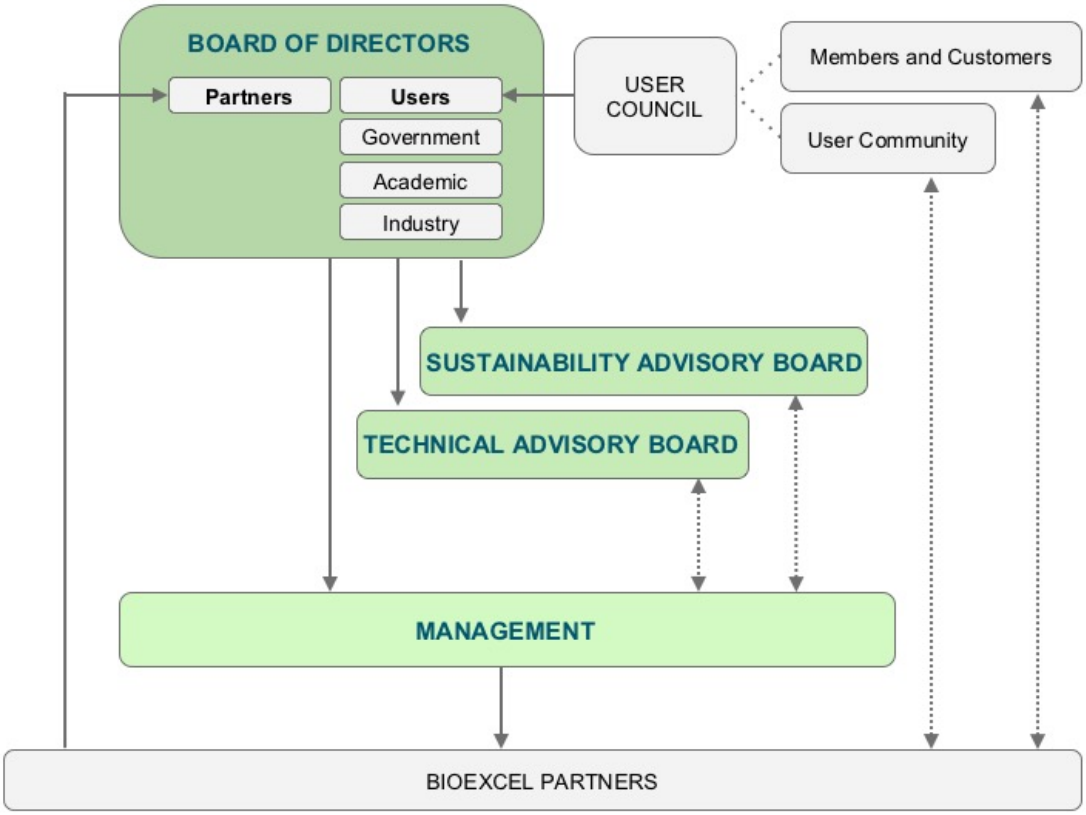
5.1 Governance

BioExcel's intention is to be user-governed, so that strategic decisions are aligned with long-term and short-term needs of our users. It is thus imperative to include representatives from the biomolecular research community in the board of directors.

The evaluation of the impact of BioExcel's work should include feedback from independent parties, and should focus on the quality of computational tools provided and the scientific output that these enable. This can be done by setting

up advisory boards with representative stakeholders, who support the board of directors and provide input to the management of the CoE.

Taking these requirements into account, the goal is to set up a slim, cost-effective structure that allows fast decision-making processes. The scheme below gives an overview over the main governing organs and how they are interconnected.



Scheme: Proposed governance structure of BioExcel.

The four central bodies in BioExcel's governance structure are the Board of Directors, two Advisory Boards, and the Management Board. A description of all parts follows below.

User Council: BioExcel will ask users from its community, as well as paying users to suggest representatives to join a virtual user council. Once a year the user council elects user-representatives into the board of directors. It is important to have user representatives from government, academia and industry are represented the board. In practice, this could be done by generating a list with candidates that represent each of these three segments. Candidates in this list are suggested by members of the user council and members of the current board of directors.

Board of directors: The board consists of the user representatives elected by the user council, and partner representatives, elected by the BioExcel partners. The number of members should be kept small to allow efficient decision making processes. A suggestion is to have a total of 4 user-representatives and 4

BioExcel partners, i.e. in total 8 board members. The board should have annual meetings to discuss and decide on the center's strategy, and at least one meeting together with the management to follow up on progress. The board appoints members for the two advisory boards.

Technical advisory board: Members of the technical advisory board represent stakeholders from the e-infrastructure and software communities with expertise on system software, data privacy, and hardware.

Sustainability advisory board: Members of the sustainability advisory board represent experts in business development, innovation management and industry standards.

Management: The management team should be kept small, e.g. depending on the centre's workload 2-3 persons. Its duties are to oversee and coordinate the work of the BioExcel partners, to ensure that the work aligns with the strategic goals, and to act as contact point to paying users. It will furthermore deal with accounting, legal contracts, dissemination and monitoring the impact of support service offerings.

Partners: Initially the partners will consist of the current BioExcel partners. However, the composition should be kept flexible, allowing partners to leave or to take in new competences if it is deemed necessary to properly deal with software development and user needs.

5.2 Legal structure

BioExcel can be a legal entity of its own, it could be set up as consortium of distributed partners, or it could choose to partner with an already established organisation. To understand the benefits and disadvantages of these options, we need to compare them in regards to their:

1. Ability to implement the chosen governance model
2. Suitability to accept different types of funding streams
3. Cost of maintenance, i.e. legal and administrative burden
4. Suitability to deal with policies of partner organisations for 'external' engagement of their employees

Partnering: Partnering with an existing organisation would dilute this organization's existing funding streams. To have an attractive proposition when approaching an organisation with complementary service offerings, BioExcel would need to bring its own funding and user base into that organisation, and clearly identify the synergistic benefits to leverage on for the benefit of the combined user communities.

Setting up a non-profit: Most service providing organisations similar to BioExcel are set up as non-profit, i.e. foundations or charities. The main benefits are reduced cost due to tax reductions. This approach works fine if the main revenue streams stem from membership fees, donations and funding. Complication arise

when charging service fees, as could be the case when interacting with industrial users. In the UK it is possible that a charity owns a limited company for this purpose, where the net profits from the business are then transferred to the charity. However, this requires setting up two organisations and keeping two annual accounts etc.

We are looking at and communicate with similar organisation who have transitioned from public funding to a sustainable model such as Open PHACTS, PRACE, or FAIRDOM. In parallel we have also started to look into the national rules for setting up foundations or businesses to get a first comparison of cost and liabilities. The table below is by no means exhaustive and only provides a first snapshot.

UK			
Structure	**Tax and VAT**	**Start-up**	**Declaration rules**
Charitable company	• 20% VAT on all standard-rated goods • National insurance rate 13,8% • Exemptions from corporate tax, capital gains tax, inheritance tax on gifts	3 founders	Submission of annual reports and accounts to Companies House **AND** Charity Commission
Private company limited by guarantee	• 20% VAT on all standard-rated goods • National Insurance rate 13,8% • 20% corporate tax • 18-28% capital gains tax • Inheritance tax on gifts	3 founders	
SWEDEN			
Ideell förening (non-profit association)	• 25% VAT on all standard-rated goods • 31% Social fees • 22% corporation tax • 30% capital gains tax	3 founders	Declaration needed if value of assets exceeds SEK1.5 million, or if the organisation conducts commercial business.
Allmännyttig förening (charitable association)	• 25% VAT on all standard-rated goods • 31% Social fees • Exemptions from corporate tax, capital gains tax, donation tax • Taxed for real estate and operating income, unless operating income is to 70-80% derived from activities that either have natural link to the association's public purposes or used as a source of funding for voluntary work.	3 founders	Declaration needed if value of assets exceeds SEK1.5 million, or if the organisation conducts commercial business.

Ekonomisk förening (economic association)	• 25% VAT on all standard-rated goods • 31% Social fees • 22% corporate tax • 30% capital gains tax	3 founders	
Aktiebolag (limited company)	• 25% VAT on all standard-rated goods • 31% Social fees • 22% corporate tax • 30% capital gains tax	1 founder SEK 50.000	
Europe			
EEIG (European Economic interest grouping)	• Unlimited company, meaning any profit or loss it makes is attributed to its members. • Qualifies for VAT • Exemption from corporation tax	2 founders from different EU countries	

5.3 Discussion

Before expanding the list of national organisations for other countries such as e.g. the Netherlands or Germany, we need to investigate the options for collaboration with existing organisations. We have informal partnerships with Open PHACTS and FAIRDOM through our members, that could become more formalized in the future if deemed feasible. To this extend we have to understand how their service offerings overlap with ours, what the benefit for the combined user communities would be, and what the conditions for an eventual partnership are.

Another important factor to consider are the policies of BioExcel's consortium partners. We have already approached the administrative and legal departments of some BioExcel partners, and it turns out that there is a variety of rules for engagement with external organisations and provision of commercial grade services. It might be difficult to find a common ground that all partner organisations can agree upon, and it might be needed to form a core-organisation of some BioExcel partners, to which the other partners with more restricted policies are bound by consortia agreements.

6 Cost and revenues

6.1 Cost types

The different types of cost that will occur at the center can be sorted into three main categories:

1. Operational costs, mainly related to administration, management, contact to users and marketing. The number of FTEs needed scales with size of the center.

2. Support costs, related to software development and general user support. The number of FTEs scales with number of software applications and workflows that are enhanced and the size of the user community.

3. Service costs, related to the provision of premium services. The number of FTEs scales with the number of paying users.

6.2 Cost estimate I: Full scale

To improve efficiency, scalability and usability of software, deal with bug fixes, and provide support for a combination of distributions, processors, and middleware environments, we assume a realistic workload of 5 FTE for each of the main software application packages.

Attractive and highly functional software with corresponding workflows and interfaces will in turn increase the number of community users that find our support offerings useful, and hence increase the workload on general user support, and the interest in our premium services.

In this model, about two thirds of the cost are related to software and infrastructure development, and another third is related to provision of user support and services.

Operational costs	FTEs needed
Support	
Software development with 5 FTEs for each of our three main software packages, plus 5 FTEs for workflows, interfaces and web portals	20
User support including maintenance of website, regular provision of webinars and tutorials, plus free training events for non-profit users.	3
Services	
Provision of premium web portal with help desk, premium training sessions, and consulting.	8
Operations	
Administration, accounting, management, contact point to users, legal support and active marketing of our premium service offerings	3
SUM	**34**
Annual cost circa	**€4.570.000**

6.3 Cost estimate II: Intermediate scale

This cost estimate assumes limited resources for development of software packages and workflows. As a consequence, fewer community members will interact with our free support. And since the premium service offerings, such as the web portal and the training are directly depended on the quality of the available software infrastructure, these premium service offerings will be less attractive to users as compared to the full scale scenario.

	FTEs needed
Support	
Software development with 2.5 FTEs per software package, plus 2.5 FTEs for workflows and interfaces.	10
User support including maintenance of website, regular provision of webinars and tutorials, and a reduced number of training events for non-profit users.	2
Services	
Provision of premium web portal with help desk, premium training sessions, and consulting.	4
Operations	
Administration, accounting, management, contact point to users, legal support and active marketing of our premium service offerings	2
SUM	**18**
Annual cost circa	**€2.420.000**

6.4 Cost estimate III: Small scale

This cost estimate considers a "survival" scenario, assuming that public funding has ceased, so there are no resources to finance software development and provision of user support. The purpose with this model is to keep the internal structure intact, and maintain a small scale service offering until new funding is secured and full operations can be resumed.

However, since in this scenario there are no developers in the center, it will be impossible to provide consultancy services The premium services are confined to training and provision/maintenance of the premium web portal. Without updates and improvements of the underlying open-source software, enhancements in the functionality of the premium web portal, this situation is likely unable to survive longer than two years.

	FTEs needed
Support	
No software development	0
Limited user support including maintenance of website and discussion forum	0,5
Services	
Provision of premium training, premium web portal with help desk	3
Operations	
Administration, accounting, internal coordination, management, contact to users, board meetings	0,5
SUM	**4**
Annual cost circa	**€538.000**

In this scenario, revenues from premium services must be sufficient to cover the entire costs of the center. This means that service fees must be set at a profit margin of 25% or higher.

6.5 Pricing of services

To estimate the service fees and membership fees that we could propose to users interested in our premium services, we make the following assumptions:

- Premium web portal: Assuming the free web portals mentioned in 4.3 are up to date, another 1.0 FTE is needed to bundle these and add premium functionality, maintain the portal, and deal with help-desk requests. It is estimated that this one FTE can deal with 1000 web portal users (not all of them will use the help desk simultaneously).

- Premium training: 0.2 FTE are needed per 100 participants. This estimate is based on a training session of 5 days, with 10 participants per class, i.e. a total workload of 10 weeks per year.

- Consultancy: 0.5 FTE are needed per 10 consultancy assignments, if the average duration per assignment is two weeks.

- The average cost of salary in BioExcel is €7000 per person month. We calculate with 60% overhead, which increases the cost per FTE to €11.200.

The table below shows our self-cost per paying user and different prices for margins between 20-50%.

Service fees for different price margins	Self-cost €/user	20% margin €/user	30% margin €/user	40% margin €/user	50% margin €/user
Web Portal + Help Desk	134	168	192	224	269
Training	269	336	384	448	538
Consultancy 2-week assignment	6720	8400	9600	11200	13400
Consultancy per hour	84	105	120	140	168

Table: Estimated prices per individual user for our premium service offerings with different profit margins.

Based on the considerations made for the small scale cost scenario, we would need to set our prices above a 25% margin to cover our self-cost plus some support and basic operations of the center.

User from non-profit organisations have already access to our free training provided as part of general user support, and can also use the existing free web portals. Based on the comparably low cost for the premium web portal with significant benefits in customization and help desk, it might be likely that the majority of users is interested in this offering. Assuming that the ratio between non-profit and industrial users is 9:1 for the web portal and training offerings, we could test following pricing with our users:

Proposed pricing offering	Average price per user	Non-profit user	Industrial user
Web portal + helpdesk, circa 25% margin	€180/year	€120	€700
Training, circa 35% margin	€400/year	€250	€1750
Consultancy, circa 45% margin	€12.000/2-week	€150/h	€150/h

Table: Proposed prices that we could test with our users for our premium services.

6.6 Revenue estimates

BioExcel will be financed through a mix of public funding, membership fees, and service fees. Public funding will cover improvement of efficiency and scalability of freely available open-source software and basic user support. Membership and premium service fees are intended to cover the costs of our premium services, and to some extend basic operations and free user support.

The revenue scenarios below are based on the pricing suggested in section 6.5, taking into account increasing demand for premium services and users support with improving quality of the underlying software applications.

	Full scale	**Half scale**	**Small scale**
FTE: Software development	20	10	0
FTE: User support	3	2	0.5
FTE: Operations	2	2	0.5
FTE: Premium services	8	4	3
TOTAL FTE	**34**	**18**	**4**
Annual cost [€]	-4.570.000	-2.420.000	-538.000
Web portal + help desk: # of users	6400	3200	2700
€180 per user/year	1.152.000	576.000	486.000
Training: # of users	400	200	150
€400 per user/year	160.000	80.000	90.000
Consulting: # of 2-week assignments	16	8	0
€12.000 per assignment	192.000	96.000	0
Premium service revenues [€]	**1.504.000**	**752.000**	**546.000**
Public funding [€]	**3.066.000**	**1.667.000**	**0**

7 Sustainability objectives

Sustainability objectives presented in this section emerge from the initially defined success goals, and take into account findings from the market research and the specifics of our value proposition. Each of these sustainability objectives should be covered by one of our current work packages. The overview presented in this section serves thus as reference to validate that our ongoing efforts are aligned with the sustainability success goals. Exemplary performance indicators are suggested for each objective, and it is up to the work package leaders to decide if these are to be included in the list of key performance indicators.

7.1 Objectives

Objective A1O1: Users can stem from academia, government and industry. Our objective is that each of these segments should be represented with more than 10% in our user community. A WP3 performance indicator could be the percentage of community members steaming from academia, government or industry.

Objective A1O2: Users can be experienced or entry-level in regards to their use of parallel computing techniques. To ensure that we are not neglecting one or the other, each of these two user types should be represented with more than 20% in our user community. A WP3 performance indicator could be the percentage of community members that can be assigned a labelling of being experienced or entry level.

Objective A1O3: Our main software application packages can be used in different biomolecular research fields. For a sufficient spread throughout biomolecular research fields, we should aim to have several, clearly distinct biomolecular research fields represented in our user community. A WP3 performance indicator could be the number of distinct biomolecular research areas within our user community.

Objective A2O1: At least 2% of our community members engages with us, either during webinars, the discussion forum, or the software feedback tool. Assuming a community size of 10.000 members, this should correspond to a minimum of 200 interactions per year. WP3 and WP4 could use the number or percentage of active users as performance indicator.

Objective B1O1: Improve the efficiency, scalability and usability of our supported software packages. The software development, and judgment of progress towards technical performance goals is part of WP1 and WP2. A possible objective that could be added here is engagement/collaboration with the Performance Optimization and Productivity Center of Excellence (pop-coe.eu).

Objective B1O2: Agree upon and implement policies for documentation of the software development process, e.g. versioning reports, traceability etc. Such documentation is important to enable independent co-development by the community, other academic actors or ISVs. A possible performance indicator for WP1 and WP2 is the number of common policies that our software developers adhere to.

Objective B1O3: Provide manuals or tutorials for the majority of newly developed software features. Such documentation is important to improve usability of software. The percentage of documentation for newly developed features could be suitable performance indicators for WP1 and WP2.

Objective B2O1: Develop workflows that increase efficiency and usability of biomolecular parallel computing software. The workflow development, and judgment of progress towards technical performance goals is part of WP2.

Objective B2O2: Develop interfaces and platforms that increase efficiency and usability of biomolecular parallel computing software. This objective plays into WP1 and WP2.

Objective B3O1: Provide and interface where users can leave feedback on our software applications, and stimulate users to provide feedback through platform. A possible performance indicator could be the number of feedback received by software programmers.

Objective C1O1: Provide platform for discussion among users and experts. A possible performance indicator is the number of new posts per month or topic.

Objective C2O1: Support community in finding matching expertise or potential collaboration partners. The number of matches/collaboration opportunities suggested to users could be a performance indicator.

Objective D1O1: Provide webinars relevant to user community. This belongs to work package 3, and the number of attendees would be suitable performance indicator.

Objective D1O2: Provide workshops and similar training events to user community. This objective belongs to work package 4, where the number of events per year, or the number of attendees per year would be suitable performance indicators.

Objective D2O1: Provide enterprise quality services to industrial users. This objective belongs mainly to work package 3, but requires input from WP1 and WP2. A suitable performance indicator could be the number of test cases/collaborations with industrial users.

E1 Work satisfaction: At this point we have too little knowledge to identify clear objectives, and further research is required.

Objective E2O1: Provide opportunities for BioExcel experts to increase their professional skills. This would objective could belong to WP6, where the number of relevant educational or knowledge sharing events pointed out to BioExcel partners would be a suitable performance indicator.

Objective E3O1: Initiate new collaboration projects with participation of BioExcel experts. A possible performance indicator is the number of joint-collaborative applications submitted.

Objective F1O1: Develop a professional graphical appearance of our website. Possible performance indicators are the number and durations of website visits.

Objective F1O2: Achieve a high visibility of BioExcel in the community and among stakeholders. A possible performance indicator is the number of events where BioExcel is formally represented with a stand or a poster.

Objective F2O1: Membership and service fees should contribute to our funding needs. A possible performance indicator could be the number of users that agree on our pricing proposal.

Objective F2O2: Identification of organisational structures suited for our governance model and source of revenue. A possible performance indicator is the number of suitable organisation types identified.

7.2 Supportive tasks

At the end of this section, supportive activities are listed that need to be worked on for the next revision of this sustainability plan in order to validate assumptions, refine objectives, and provide references for the alignment of key performance indicators.

User related:
1. Identify top 3 biomolecular research fields using HPC computing
2. Estimate number of biomolecular researchers using HPC
3. Identify user perceived challenges in use of HPC
4. Estimate number of biomolecular researchers using HPDA

Software related:
1. Define technical and usability challenges of cloud delivery of HPC services
2. Validate community interest in the proposed support offerings
3. Define approaches to measure increase software usability among users

Support related:
1. Invite users and partners to participate in discussion forum
2. Invite users to use software feedback tool
3. Compare approaches to facilitate matchmaking
4. Identify national funding opportunities suited for collaboration between BioExcel and user community

Service related:
1. Validate community interest in service propositions
2. Interview ISVs to understand requirements for co-development
3. Interview industrial users to understand service requirements

Partner related:
1. Interview partners to understand their training needs
2. Identify means to recognize excellence of experts with external stakeholders

Organisation related:
1. Identify suitable partners with overlapping communities & services
2. Identify benefits of merger for organisations and community members
3. Compare benefits/disadvantages of organisational types

Market study:
1. Investigate the e-infrastructure ecosystem beyond HPC to include data management and analytics, where distributed computing is the biggest technology trend right now.
2. Obtain estimates for the absolute numbers of biomolecular researchers working with parallel computational techniques throughout Europe.

Stakeholders:
1. These should include public funding agencies for core services. Also, relevant research alliances or partnerships could be mentioned such as ELIXIR and the Pistoia Alliance.

Appendix

The International Data Corporation(IDC) has performed two HPC market studies on behalf of the European Commission. Their most recent study is used as the main source for input on the HPC market. At the time of the IDC study the market data for the five years from 2009 to 2013 was in place, whereas all data from 2014 onwards is based on predictions. Below is a summary of some key market data with potential relevance to BioExcel.

I: Global HPC server revenues

The tables below summarize the revenue streams from HPC server sales. The label EU+ refers to the 28 member states of the European Union plus Norway and Switzerland.

Historic HPC server revenues [billion Euro]

	2009	2010	2011	2012	2013	CAGR 2009-2013
USA	3.14	3.02	3.34	3.48	3.25	0.86%
EU+	1.61	1.97	2.09	2.16	2.01	5.70%
Asian/Pacific w/o Japan	0.64	0.87	1.10	1.15	1.38	21.18%
Japan	0.59	0.41	0.56	0.90	0.48	-5.03%
TOTAL	6.22	6.56	7.42	8.00	7.42	4.51%

Predicted HPC server revenues [billion Euro]

	2014	2015	2016	2017	2018	CAGR 2013-2018
USA	3.47	3.69	3.91	4.13	4.35	6.00%
EU+	2.16	2.32	2.49	2.65	2.80	6.85%
Asian/Pacific w/o Japan	1.30	1.41	1.52	1.63	1.91	6.72%
Japan	0.51	0.54	0.57	0.60	0.63	5.59%
TOTAL	7.26	7.72	8.26	9.33	10.13	6.42%

II: EU HPC server segments

HPC systems are used by different types of organisations and for a wide range of applications. HPC server buyers can be segmented based on common characteristics of organisations type or performance of similar activities, e.g. geosciences, financial services or weather forecast.

In their analysis, IDC did mix economic and application segments. The definitions for the four largest segments are given below. Please observe that the segment of biosciences per definition used by IDC only includes HPC systems specifically

targeted for bioscience applications. HPC systems purchased for broader scientific use that also might run some bioscience applications are instead counted within economic segments, e.g. academia or government.

Government: This segment encompasses government funded institutions that may combine both purely scientific research with research in areas of national priority (e.g. cancer research or defence). These centers don't normally offer degree programs for students.

Academic: This segment includes public or private institutes of higher education that perform scientific research, engineering R&D efforts, and educational activities. Privately funded or non-profit research institutes that work to extend the bounds of public knowledge are also included in this segment.

CAE: Computer-aided engineering and mechanical design includes disciplines such as finite element modelling, computational fluid dynamics or solid modelling. Like CAD applications, these CAE tasks are used to design automobiles, aircraft, running shoes, ski equipment, and other everyday items.

Biosciences: This segment includes use of HPC systems for genomics, proteomics, pharmacogenomics, pharmaceutical research, bioinformatics, drug discovery, bioanalytics, and agricultural research. Computational techniques include database searching and management, molecular modelling, and computational chemistry.

Historic EU+ HPC market by segment [billion Euro]

	2009	2010	2011	2012	2013	CAGR 2009-2013
Government	0.41	0.52	0.56	0.59	0.55	7.62%
Academic	0.31	0.37	0.39	0.40	0.40	6.58 %
CAE	0.19	0.24	0.26	0.28	0.25	7.10%
Biosciences	0.18	0.23	0.23	0.24	0.22	5.14%
TOTAL	1.61	1.97	2.09	2.16	2.01	5.70%

Predicted EU+ HPC market by segment [billion Euro]

	2014	2015	2016	2017	2018	CAGR 2013-2018
Government	0.59	0.65	0.68	0.73	0.75	6.40 %
Academic	0.42	0.45	0.49	0.52	0.58	7.71%
CAE	0.28	0.30	0.33	0.36	0.37	8.16%
Biosciences	0.23	0.24	0.26	0.26	0.28	4.94 %
TOTAL	2.16	2.32	2.49	2.65	2.80	6.85%

III: Location of top HPC supercomputers

Location of supercomputers that are among the world's 50 most powerful.

# of supercomputers	2012	2013	2014	2015	2016
EU+	17	16	18	18	18
U.S.	20	18	19	21	17
China	5	5	5	4	4
Japan	6	7	5	3	6

Source: top500.org

IV: EU supercomputing centers

More detailed information about the below listed organisations, their activities and budgets can be found in [IDC_HPC_2015].

Germany: Gauss Center for Supercomputing (GCS) is an alliance of three national HPC centers: HLRS (Stuttgart), LRZ (Munich) and FZJ (Jülich). GCS represents Germany in the PRACE alliance and provides three of the current six Tier-0 systems.

France: France has two national supercomputing centers, CEA and CINES. However, the Grand Equipment National de Calcul Intensif (GENCI) has the central role in HPC in France. Also noteworthy is Teratec, an association which unites over eighty technological and industrial companies, laboratories, research centers, and universities who want to combine their resources in simulation and high-performance computing. The Curie supercomputer, owned by GENCI and operated by CEA, is the first French Tier0 system open to scientists through the French participation in the PRACE research infrastructure.

United Kingdom: The UK has no permanent national supercomputing center. Instead, major centers compete periodically for the contract to provide the HPC national academic service across the UK. At present, the Edinburgh Parallel Computing Center has that role (ARCHER). The Hartree Center, which is part of the Science and Technology Facilities Council, has a major role in supporting the HPC needs of businesses in the UK.

The Netherlands: SURFsara is the national supercomputing and e-science support center for all Dutch universities, a number of large research, educational and government institutions, and the business community. SURFsara has been a partner in large European e-Infrastructure projects including PRACE, EGI.InSPIRE and EUDAT.

Spain: BSC-CNS (Barcelona Supercomputing Center – Centro Nacional de Supercomputación) is the national supercomputing facility in Spain and hosts the MareNostrum supercomputer. The mission of BSC-CNS is to investigate, develop and manage information technology in order to facilitate scientific

progress. The Spanish Supercomputing Network links MareNostrum to more than a dozen smaller HPC sites in Spain. BCS is a PRACE tier-0 host member.

Italy: CINECA is Italy's national supercomputing center and the country's PRACE host site. CINECA's Fermi supercomputer is one of the world's most powerful. CINECA is a non-profit consortium made up of 70 Italian universities, four Italian Research Institutions and the Italian Ministry of Education.

Finland: CSC, the Finnish IT Center for Science, is Finland's national supercomputing center and supports both science and industry. CSC supports a European-wide customer base of thousands of researchers in disciplines such as biosciences, linguistics, chemistry and mathematical modelling.

Denmark (minor): The Danish Center for Scientific Computing (DCSC) organizes access to supercomputing resources for Danish scientists.

Norway: The Norwegian Metacenter for Computational Sciences oversees time allocation for Norway's four supercomputing centers, NTNU in Trondheim, the University of Bergen, the University of Tromsoe, and the University of Oslo.

Sweden: Sweden has no single national supercomputing center. The Swedish National Infrastructure for Computing (SNIC) is a distributed infrastructure that is funded in part by the Swedish Research Council (Vetenskapsrådet) and in part by the participating universities. In October 2014, KTH installed a 2PF supercomputer, the largest to that date in the Nordic countries.

Switzerland: The Swiss National Supercomputing Center is CSCS (Centro Svizzero di Calcolo Scientifico), an autonomous unit of the ETH Zurich.

V: References

IDC_HPC_2010: A Strategic Agenda for European Leadership in Supercomputing: HPC 2020 – IDC Final Report of the HPC Study for the DG Information Society of the European Commission, 2010.

IDC_HPC_2015: SMART 2014/0021 - High Performance Computing in the EU: Progress on the Implementation of the European HPC Strategy, Luxembourg, Publications Office of the European Union, 2015, doi:10.2759/034719

Solve_2014: Solve, The Exascale Effect: The Benefits of Supercomputing Investment for U.S. Industry, Council on Competitiveness, Washington, 2014.

IS360R_2015: Addison Snell, HPC Trends for 2015, Intersect360 Research, online presentation.

IS360R_2011: High Performance Computing Forecast for 2011 through 2015: Economic Sectors and Vertical Markets, Intersect360 Research, 2011.

EIRG_2012: Policy Paper on Scientific Software, e-IRG Task Force on Scientific Software, The Hague, 2012.