SNIC Review

Report by Appointed Review Committee to the Swedish Research Council

Swedish Research Council

2021

SNIC Review

Report by Appointed Review Committee
to the Swedish Research Council

Dnr 2020-06593

Swedish Research Council
Vetenskapsrådet
Box 1035
SE-101 38 Stockholm, Sweden

Content

Preface

The Swedish Research Council (Vetenskapsrådet) is a governmental agency with the responsibility to support basic research of the highest scientific quality in all academic disciplines. It is also part of the agency´s remit to evaluate the research it funds and to assess the academic quality and impact of the research. The Council for Research Infrastructure (RFI) at the Swedish Research Council has the overall responsibility to ensure that Swedish scientists have access to research infrastructures of the highest quality. Specifically, RFI regularly assesses the needs for research infrastructures, launches calls and evaluates applications, participates in international collaborations and works on monitoring and assessments. SNIC is the Swedish National Infrastructure for Computing, a distributed research infrastructure for which the Swedish Research Council/RFI is the largest funder.

In 2020 RFI initiated a review of SNIC. The review was carried out during the autumn of 2020 by an international panel consisting of four experts: Nuria Lopez, Steven Newhouse, Minna Palmroth, and Thomas Schulthess. RFI wants to emphasize the importance of the panel's conclusions and especially highlight the need for SNIC to:

- Develop a national strategy with clear goals, taking into account the international development.
- Adjust the organizational structure and strengthening the executive power of SNIC management, enabling formulation and implementation of the strategy.
- Consolidate its e-infrastructures into one hardware center that may or may not be located in Sweden, while developing the local structure for software and support.
- Organize user communities in order to guarantee that strategic goals and provision of services match the needs of the research community.

RFI would like to take the opportunity to thank the review committee for their commitment and excellent work, which has resulted in this report. Furthermore, the efforts of the management and staff of SNIC as well as of the members of the user community, are highly appreciated, both in the preparation of the background material for the review, and for being available for presentations, discussions and in-depth interviews.

Irene Wennemo
Ordförande RFI

Björn Halleröd
Huvudsekreterare

1. Introduction

This is the report corresponding to the first of two reviews of the Swedish National Infrastructure for Computing (SNIC) requested by the Council for Research Infrastructures with the aim of assessing the organization in its transformation from a collaboration between six computing centers (SNIC 1.0) to a consortium of ten major Swedish research universities. The purpose of the report is to support SNIC leadership in developing its strategy for the future funding cycle starting in 2023. It is in this context that the recommendations of the present report should be understood.

In a second review, to be conducted in April 2021, the SNIC leadership will present its revised strategic plan considering the recommendations of this report. Based on the recommendations of this second review and further communication with SNIC the Council for Research Infrastructures (RFI) at the Swedish Research Council, for the continued funding of SNIC beyond 2022.

The first review was conducted via videoconference on November 27, 2020, by a panel of four members:

- Prof. Dr. Núria López, Institute of Chemical Research Catalonia, ICIQ, ES
- Dr. Steven Newhouse, European Bioinformatics Institute, EMBL, UK
- Prof. Dr. Minna Palmroth, University of Helsinki, FI
- Prof. Dr. Thomas Schulthess, Swiss National Supercomputing Center, ETH Zurich, CH (Chair)

The material made available to the panelists is summarized in the Appendix. The review itself consisted of a presentation and discussion with the SNIC Chair (Prof. Dr. Ingela Nyström), SNIC Vice Chair (Prof. Dr. Per Dannetun), SNIC Director (Prof. Dr. Hans Karlsson) and the coordinator of the SNIC Office (Per-Olov Hammargren); as well as multiple interviews with the leadership of six SNIC partners that host computing centers, non-hosting members, and two users. Furthermore, written input was requested from all SNIC partners, selected users, as well as other research infrastructure that rely on SNIC.

2. Executive summary and SWOT analysis

SNIC is a partnership of ten Swedish research universities that is mandated by the Swedish Research Council to provide national services for computing, data storage, as well as curation and management of active data sets. It does so with contributions from six data centers that are owned and operated locally by SNIC partners. The General Assembly (GA) of SNIC partners limits the scope of SNIC to services that can only be provided at a national level or reach a unique cost-efficiency. While this is consistent with its charge, the review panel concludes that this scope is too limiting. Furthermore, it was concluded that the GA is inhibiting SNIC from developing an appropriate vision and implementation strategy for a national computing and data infrastructure that is internationally competitive.

Given the limitation in scope, the SNIC director is doing a remarkable work in administering the partnership. He does not have executive power; nevertheless, he is supported by the SNIC Board, which does have executive power. However, the Board is not in charge of the SNIC strategy either. Despite that, given the limitations of governance, the decision-making process is transparent and efficient.

The main project during the current funding period is the consolidation of the transformation towards the partnership of ten universities that includes, but is no longer limited to, the six computing centers of SNIC. This transformation is clearly slipping and is perceived as an administrative measure rather than a strategic opportunity.

The review and allocation process of users' projects is operated well for conventional computing requirements, and the monitoring of resource usage is executed at a high standard. However, the panel noted that there is no national user program. Furthermore, there is no national technology roadmap either. Rather the roadmaps are set by the participating centers and user support is managed by the center locally. Traditional HPC users with conventional computational needs are well served; however, the needs of several user communities (e.g. climate) are not met. There is only a local technology outlook at the individual center level, which the panel considers to be subcritical in particular when compared with international competition. Consequently, users are not sufficiently supported by SNIC in the adoption of new computing architectures. High-end users are encouraged to find support elsewhere.

The overall findings of the panel are best summarized in the SWOT analysis at the end of this section. The panel makes numerous recommendations in the following three areas:

1. Recommendation to change the strategy process: the GA needs to broaden the scope of SNIC, properly taking into account the perspectives of users and domain science communities; users should be organized at the national level and represented in the strategy process; an appointed person or a small group with executive powers should be put in charge of developing the vision for SNIC, as well as the strategy for implementing it; the SNIC partners need to commit to this vision and strategy, and appoint the executive bodies accordingly.
2. Technical recommendations: SNIC should consider separating the underlying e-infrastructures from data platforms and HPC/supercomputing platforms; e-infrastructures should be consolidated into as few as one or no hardware centers that are directly funded, and SNIC should consider working closely with external providers of computing and storage infrastructures, such as the LUMI consortium; the cost savings that result from such consolidation should be invested into software development for data and computing platforms; particular attention should be given to early adoption of new hardware architectures by users, which should include a technology watch at the national level; overall, nationally organized users' communities should be supported proactively and investments should be made into competence development and expertise of users.
3. Management recommendations: SNIC management should be service oriented with clear accountabilities; it should be put in charge of developing a vision for SNIC and implementing the strategy; this should include keeping track of cost more transparently and keeping a risk registry with explicit mitigation plans.

Strengths:

- Good and stable funding environment.
- Well established standard resource allocation process that meets the needs of competent users who have conventional HPC requirements.
- Decentralized competences adapted to local research cultures, well connected and embedded in leading universities in Sweden.

Weaknesses:

- Weak project and consortium management due to structural problems of the governance with local priorities of partnering universities dominating the agenda.
- Culture of avoiding dispute and lack of direct communication leading to slow decision making that can be futile in a context of fast international developments.

- Poor service orientation, interactions with users are not service oriented, priorities are centered on HPC providers, rather than focused on final user.
- Users' communities are not organized and have no input on strategy; communities do not see SNIC as a source for advanced users' support.
- Distributed service organization is not effective in running a consolidated national service or exploiting synergistic effects.
- Fragmented infrastructure has no economies of scale and poor cost efficiency.

Opportunities:

- The LUMI consortium.
- Strong educational and trust culture of the Nordic circle.
- Diversity of non-traditional HPC users' communities seeking IT support.
- Well organized scientific communities, world scientific leaderships, and centers that can develop software/data platforms and perform technology watch/orientation.
- Focus on providing higher value to users through software and data platforms, rather than running hardware services.

Threats:

- High cost, which exposes SNIC to the threat of emerging public cloud platforms that will make SNIC's services look narrow, expensive and mis-aligned with users' needs.
- Cost of hosting large-systems will dominate budget at the expense of equally or more valuable services to users (e.g. training, advanced user support).
- International visibility is poor.
- End of Moore's Law and Cambrian explosion of architectures, as well as inability to respond to architecture diversity.
- Mismatched resources and outcomes, which could inflict a severe damage to the scientific community in HPC.

3. General impression based on discussions with SNIC leadership and partners

3.1 Charge and scope

The charge SNIC received from the Swedish Research Council is to provide a set of national services for large-scale computing, as well as data storage, curation and management of active data sets; along with a specific set of dedicated national services for supporting other research infrastructures; as well as a coordinated effort on advanced users' support.

The scope of SNIC is defined by the General Assembly of SNIC partners and is consistent with this charge. However, the scope is limited to services that (1) can only be provided at a national level, or (2) reach a unique quality and cost-efficiency. There have not been any enhancements or changes in scope, despite strong evidence based on users' testimonies that the needs of users outside a few traditional, compute-dominated domains (e.g. chemistry, physics, biophysics and fluid dynamics) are not well supported. Where enhancements have started, they have been slow to progress (e.g. GPUs, Cloud) and there is a danger SNIC will miss these opportunities and user communities will look elsewhere.

While the panel considers the limited interpretation of the scope of SNIC's mandate to be the root cause for most difficulties and delays experienced by SNIC, we will discuss the present cost and schedule, as well as all technical and managerial aspects in the context of the present scope. We will come back to these limitations in the conclusion and recommendations below.

3.2 Cost and schedule

The cost of compute services is relatively high (50% higher than expected for large data center infrastructures) but this can be explained by the fragmentation of the infrastructure over six compute centers that each run independent procurements, none of which reaching sufficient size to ensure economies of scale. Moreover, the cost of storage (dCache and IRODS) is prohibitive (2-5 times compared to other, similar infrastructures in Europe). This cannot be explained with fragmentation and requires further analysis.

While the local supervision of SNIC staff does in general strengthen the organization's embedding in the university system, the coordination of skills at the national level constitutes a challenge. Furthermore, it has happened that local

priorities are ranked higher than those of SNIC. This can impact cost efficiency and robustness of the user support and manifests a weakness of the in-kind contributions system. However, these deficiencies are recognized and the monthly reporting of activities at the partner sites provides adequate tracking of the implementation of in-kind contributions.

The transformation from SNIC 1.0 to 2.0 is clearly slipping and runs behind schedule. From the discussions it is evident that some of the centers have not yet embraced the change to SNIC 2.0. The transformation is often perceived as an administrative measure rather than a strategic opportunity. This is probably a manifestation of the limited scope and missing vision of SNIC. This also shows in the statement that experienced users would be redirected to EuroHPC, if SNIC funding were to be reduced. Redirecting users is an administrative measure that does not give justice to EuroHPC as an opportunity.

3.3 Technical

There is only limited coordination of the technical roadmap at the SNIC level. The processes are standard and uninspired: analysis of usage, dialog with users is mostly limited to local interactions (no users committee exists), and local technology outlook. Procurements are managed by individual partners/centers based on their own roadmaps. There is no integrated technology watch nor strategy at the SNIC level.

Several partners/centers made clear that SNIC is just one of many priorities, a so-called national layer. There seems to be no real incentive to change established roadmaps. This is unfortunate, since SNIC loses the synergistic potential and economies of scale it could have. Furthermore, user communities that are not directly attached to one of the center partners have no opportunities to provide input on the roadmaps. It appears that the SNIC center partners do not embrace community specific requirements, as the above-mentioned example with the climate community shows.

Consistent with the strong local autonomy, there is no risk registry at the SNIC level. Risks are managed locally by the centers.

3.4 Management

Overall, given the limitations in scope, the SNIC director is administering SNIC as best he can. He does not have executive power and thus is limited in what he can direct. His role would be best defined as that of a managing director, rather than that of a director in charge of a long-term vision.

The Board has executive power; however, proposing a strategy does not fall under its responsibility either. The General Assembly is responsible for and

decides on the strategy but again, the local competences and focus represented there steer the outcomes of the discussions. With no users' community stakeholders represented in the General Assembly there is a lack of diverse thinking and viewpoints.

Considering this rather peculiar strategy approach, the decision-making process appears to be remarkably transparent and efficient. However, the setup is such that SNIC cannot develop a vision, in stark contrast to competing structures all over Europe. The strategy process is designed to emphasize the local interest of partners and is thus bound to converge to the least common denominator. SNIC is not in a position to take a leadership role in the country and the users or scientific communities are limited to a passive role.

4. Feedback from users

In general, it is to be noted that Sweden has users of computing and data infrastructures that are world-class researchers in their own fields. For users, SNIC presents itself as a source of computing resources that are made available through open calls every six months for yearly allocation periods. The process is such that the user sends in a short application, which is peer-reviewed by members of the Swedish National Infrastructure Allocations Committee (SNAC), with support of external reviewers. The review process is of high quality, and it is to be noted that the tool with which the previous usage of the user is monitored, is exceptionally performant.

The panel is slightly concerned about the work effort that the users carry out, given the relatively small grants they get. For example, in other countries, grants below 100,000 node hours p.a. (~4 million CPU core-hours on older machines and ~10 million on CPU core-hours on newer systems) can be given with a simple technical review, while the peer-review process is applied only when the proposal is over this limit. Furthermore, for some users the allocation schedule is not optimal, given the boundary conditions in the fields they operate with high localized demand peaks that cannot be accommodated.

The local expertise that the users receive from their local centers is mainly conventional HPC expertise. There is a strong demand to grant the users' needs at all levels: (i) education; (ii) feedback from using resources; (iii) future advances and challenging new architectures, (iv) regular feedback from the users' community to the normal operation of SNIC. The local centers do not provide foresight into the emerging architectures, and therefore the users who are currently thinking of the emerging architectures, are left on their own.

5. Discussion and recommendations

The SNIC partners defined the scope of the organization as a “coherent national e-infrastructure supporting all areas of Swedish research that have needs of large-scale computing and/or large-scale data storage/management of active data sets.” SNIC services are to be defined in such a way, that they can only be provided at a national level, or such that they provide the best and most cost-efficient support to researchers. The partners appear to be driving the SNIC activities, which are heavily dominated by the independent strategies of the participating local computing centers. While the computing services mostly meet the needs of the established HPC communities, SNIC overall portfolio is not meeting all needs of the communities they are commissioned to serve. E.g. SNIC is unable to meet the needs of the climate community with regard to data storage and support of their workflows. Similarly, while the (well-known and established) needs of the World-wide LHC Computing Grid (WLCG) are properly met mainly through the engagement of WLCG focused SNIC centers, the SNIC services quality delivered to smaller high-energy physics projects is perceived to be lower.

Limitation in the scope and the role of the General Assembly to inhibit SNIC from developing a clear vision, seems consistent with the history of the partnership and cultural peculiarities of traditional HPC centers. It started with three university-based computing centers in 2003, and was later expanded to six centers. Given this background, handing over control to SNIC of ten partnering universities in SNIC 2.0 will not automatically lead to a change in priorities. The strong control of the centers persists, and it is quite natural that the partners want to limit the role of SNIC to national affairs, in order to protect local priorities. Participation in SNIC gives the partners access to stable national funding without sacrificing autonomy to a national competition. The panel does consider the partnering model with strong embedding in the research university systems a strength; nevertheless, control of the strategic agenda should be moved from centers to researchers with appropriate changes to the governance structure.

The biggest impact of the conservative strategy can be seen in the user program. While well executed within the given scope, the user program is clearly missing several aspects of modern extreme-scale computing and data infrastructures:

By biasing new allocations on carefully monitored previous usage of the respective research groups (bias on past trajectories), incremental evolution of resource access is unavoidable, which will in turn prevent users from proposing challenging new projects. Large users are encouraged to go elsewhere, which is not a good strategy to develop a high standard for extreme computing within

Sweden. In contrast, the SNIC strategy seems to keep users where they are, as exemplified by the often untransparent communication of outcomes from the proposal reviewing process, by not providing proposers with the information needed to improve subsequent submissions.

Fragmentation leads to high (compute and data) infrastructure costs, as well as missing synergies in SNIC. There is no technology outlook at the national level, and any local technology developments are subcritical and often too late. Sweden will miss out on the rapid developments of computing and data technologies. Native-cloud technologies applied to supercomputing and scientific data platforms, which lead to cost efficient high-quality services that are very adaptable to customers' needs, are being developed elsewhere.

With the end of Dennard scaling and Moore's Laws there has been a Cambrian explosion in computing architectures that is hard, if not impossible to keep track of at the level of local HPC centers. The resulting missing visibility of architectural developments in the users' communities has consequences on applications, since software developments are falling behind and adaptation to new architectures are simply missing.

Consequently, Swedish scientists will have to go elsewhere to find what they need to remain competitive at a global level, and SNIC will lose its national relevance. That will be in contradiction to SNIC's scope, which is limited to the national level. This contradiction is a result of the current strategy process, by which SNIC is bound to converge on the least common denominator between partners.

5.1 Recommendations to change the strategy process

The review panel is convinced that the strategy process is deficient and could threaten SNIC's future. The recommendations will thus start at the top, with governance and the need for SNIC to develop a strategic vision.

- The General Assembly is encouraged to broaden the scope of SNIC, properly taking into account the perspectives of users and domain science communities.
- For this, the user and domain science communities should be organized at the national level and represented in the strategy process, along with the centers.
- An appointed person, or a small group with executive powers (e.g. the SNIC Board), should be made responsible for developing the vision for SNIC, as well as the strategy to implement the vision.
- The SNIC partners need to commit to the mission and strategy of SNIC, and appoint the Board/executive body accordingly.

5.2 Technical recommendations

There are several technical opportunities related to the emerging trends in cloud-native computing for science. First and foremost,

- SNIC should consider separating the underlying e-infrastructures from data platforms and HPC/supercomputing platforms.

On one hand, the e-infrastructures provide bare computing and storage services at the lowest possible cost. This will only be possible if SNIC works with e-infrastructures that are large enough to reach sufficient economies of scale. Thus,

- SNIC should consolidate its e-infrastructures to as few as one or no hardware centers that are directly funded by SNIC.
- SNIC should consider working closely with external providers, such as the LUMI consortium where resources could be acquired/hosted in Kajaani.
- SNIC partners could consider moving their local infrastructure operations to virtual clusters running on these remote e-infrastructure.

Given the high potential cost savings through the separation of e-infrastructures from (HPC) compute and data platforms, SNIC should be able to invest the balance of the funding into the development and operations of precisely these platforms. If about ⅓ of the current hardware costs are saved, approximately 50 million SEK could be available annually for investments in software development, competence building in the users' community, as well as exploration of new fields. Specifically,

- SNIC should consider establishing a program that funds the development and future operation of computing and data platforms;
- These platforms can be domain specific and organized around large software packages, where international collaborations (e.g. LUMI/EuroHPC, EOSC), as well as collaborations with external data platforms should be sought (not all software should be developed/maintained in Sweden); and
- SNIC could consider investing part of its budget in the exploration of new fields.

Such a program would directly strengthen the Swedish users' communities at a national level. In any case, SNIC should

- Invest into competence and expertise of the users;
- Engage and organize the users' community and scale it; and
- This support should be proactive.

In response to the end of Moore's and Dennard's Scaling laws with the implied explosion in architectural diversity,

- SNIC should introduce a technology outlook at the national level; as well as
- A SNIC strategy to more aggressively adopt emerging architectures; and

- In particular, SNIC should encourage support staff and application scientists to rapidly adopt accelerated computing architectures, in light of LUMI, but also to foster the careers of young researchers with the adoption of the latest technologies.

Such as strategy could be developed in collaboration with computer science and engineering departments at Swedish universities.

5.3 Management recommendations

The changes in the overall governance and strategy process will have implications for the management of SNIC. The panel recommends that

- Somebody be put in charge of implementing the SNIC strategy -- this could be the SNIC Director in a new executive role or an appropriately appointed SNIC Board.
- The users' community needs to be organized, in order to facilitate the engagement of users and domain science communities in the strategy process.
- Overall, the SNIC management should be service oriented with clear accountabilities.

Implementation of a global strategy will require earned value project management with appropriate attention to scope, cost, and timeliness, as well as risks. Specifically,

- The SNIC management will have to keep track of costs more transparently;
- SNIC will need to introduce a risk registry with explicit mitigation plans; and
- An overall metric should be introduced to monitor and keep track of earned value in the development of all infrastructures (software, data and computing platforms), as well as the user program including access to infrastructures but also education.

Appendix: List of review materials

The following material was made available to the review panel:

- Terms of Reference
- SNIC funding application
- SNIC self-assessment (from Uppsala University review)
- Partner feedback (from Uppsala University review)
- Letter with questions from the panel, sent to SNIC partners
- Strategic Implementation Plan. Adapted by the SNIC board, translated to English.
- Contact information to select SNIC users (provided by SNIC)
- Expenditures for SNIC computing and storage resources. Provided by SNIC.
- Timeline for SNIC resources. Operational and planned resources.
- SNIC Annual Report 2019. Translated to English.
- Funding key. Cash and in-kind contributions per partner.
- Letter with questions from the panel to user communities and infrastructures.
- SNIC organogram
- SNIC presentation slides from review meeting
- SNIC Halftime Review final report. SNIC review commissioned by the SNIC board, administered by Uppsala University.
- SNIC Halftime Review terms of reference.
- Written answers to panel questionnaire, from SNIC partners and from user communities.

The Swedish National Infrastructure for Computing (SNIC) is a national research infrastructure that makes available large scale high performance computing resources, storage capacity, and advanced user support, for Swedish researchers. SNIC is a consortium of 10 universities, with Uppsala University as host organisation. SNIC is funded by the Swedish Research Council for the period 2018–2022. In 2020, the Council for Research Infrastructure commissioned a review of SNIC by an international panel. This report contains the panel's findings and recommendations.

Swedish Research Council
Västra Järnvägsgatan 3
Box 1035, 101 38 Stockholm, Sweden
Tel +46 (0)8-546 44 000
vetenskapsradet@vr.se
vr.se/english