I hesitate to call this response a white paper. It is a set of comments from me, Mark Hahn (hahn@sharcnet.ca) as an individual staff member. I apologize in advance for the relative lack of polish and concision.

Things ARC has to get over in Canada:

- CFI. CFI was never an appropriate funding mechanism, even if it was the only option at the time. CFI is oriented towards specific projects, rather than organic, evolving ecosystems. CFI seems to do well with projects like building hospitals, but its preferences have not served ARC researchers well.
- The HAL report. This report, required by CFI, poisoned the ARC ecosystem for a decade. It's not clear that consultants can ever offer useful insight into complex topics, no matter how much the governance types want it. Among the outright mistakes in the report was the MBA-centric obsession with minimizing the number of sites, and insisting on a "competitive" resource allocation process.
- The ecosystem is starving to death. This means that most members of the ecosystem feel threatened, and researchers can never get what they need. Almost two decades ago, the "Long Term Plan" showed that consistent, regular funding of about $50M/year was required. Nothing about the funding so far has been consistent, and the country is probably in $150M of ARC "deficit".
- Partly as an outcome of these mistakes, sites have never had incentive to cooperate. Since funding is zero-sum, and CFI treats each site as a separate project, it is natural for each site's host institution to own the hardware, and run the procurement process. This is what the HAL report should have really focused on.

We do need a way to cooperate, and/or more unified governance. We need to become more coherent than just a federation.

It's also true that we need to accommodate differences. Some of these differences are inherent. For instance, there are smallish fractions of the community who definitely want 24/7 operation. The problem is that this is expensive, and money spent providing that kind of staffing cannot be spent on the facilities that will enable other research. There are many aspects of ARC that are like this, where a property is much desired by one constituency, but clearly against the interests of another. Another example is computer memory: some subcommunities live or die based on whether they can get TBs of memory in a single host, but others never want more than a few hundred MB per core. Some researchers have no use for GPUs. There are even domains where fast interconnect and filesystems are just wasted. We need a governance structure which can equitably divide the limited funding among these conflicting desires. This has to happen by a transparent process (so constituencies can understand the choices) that puts all investment trade-offs on a single table at once.

Commercial cloud is an existential threat to shared academic computing:

Clouds can do anything. Commercial clouds can do it well, and do it fast. The only problem is that the companies selling cloud access doing it are motivated specifically to make very high profits. When Amazon rents you a server for a year, their cost is what you'd pay to own and operate it outright. So over 3 years, you pay three times as much. Any money spent on commercial cloud harms the shared-academic research effort.

We need to incentivize improvement:

RAC is a great example. It exists because CFI has an obsession with "competitive" allocation of resources. To the bureaucratic mindset, that means: collect all the asks, apply an evaluation, and fully fund the best, grudgingly provide resources to the OK ones, and cut off the losers. There are many problems with this:

- That we can pick winners. Competitive processes like this tend to devolve into beauty pageants - well-known researchers get high scores. But notoriety is different from merit! This is part of the reason that most of the innovative work that happens in CC today, occurs in “grudging” default allocations.
- Since this heavy-handed process is used to pick winners, it is impractical to run frequently, and it’s important to minimize the number of applicants. So we wind up with a process that consumes a lot of staff effort, is onerous for researchers to apply, and which only meets the timing needs of certain kinds of projects (which can plan a year ahead, and can expect to consume similar resources at a constant rate for the entire year).
- Not only is this process not effectively merit-based, but it’s also inefficient. An applicant must apply pad the amount of resources requested (to appear important), and must ask for all possible needed kinds of resources. We wind up with many projects that drastically underuse their allocations and have no good answer to the many researchers who need an immediate burst of access.

Why have we not improved the RAC process, since it has these obvious problems? I could speculate, but at the very least, CC as currently constituted is simple not able to do so. We need lightweight applications, preferably with scoring taken from TriCouncils, that can allocate resource amounts as short as a week, and which operates at least monthly. There are many technical implications of this, but it is clearly what researchers need, and what NDRIO should mandate.

There are many, many other areas where we have been insufficiently ambitious in fixing problems we know about. For the most part, this lack of ambition is directly tied to the funding and governance shortfalls.

Acquisition:

We need a unified acquisition process, so that there can be a national RFP run every year, and hardware for all facilities comes from the results. It is a fallacy that vendors care that much about big projects. They want our business, and their fulfillment systems are efficient enough to work well even on modest orders. We do not get giant discounts when we order $10M at a time. Economies of scale do exist, but only make a difference when you compare single orders to moderate-sized ones. Once you get a decent-sized order, costs become nearly linear. This means we should not fixate on large, single-install projects - it’s more important to retain agility so that we can respond to new or changing needs.

Allegiances:

We need all our staff to be 100% dedicated to the project. Divided loyalties inevitably result in contradictory incentives. At the same time, we need to avoid the rot of a top-down management style. Our staff are the strength of the organization, and decisions need to be participatory and the organization flat. Avoiding divided loyalties does not necessarily translate to having a single way of doing everything. There are times when diversity is a strength, as long as the diversity does not introduce efficiency or complexity problems. Diversity should be inherent to our decision-making processes - not unlike how GPUs are good for some things and not others. What we need to avoid is gratuitous differences where there is no case for difference except taste.

Similar to staff allegiance, we need clarity on assigning other resources. For instance, certain academic domains have traditionally had very loud voices, in part because they have long histories of involvement, in addition to outside factors. I’m thinking of ATLAS and related particle physics projects, which command dedicated personnel and hardware, and which are expected to dominate the RAC process. I am not arguing against this field. I am arguing that there should be a level playing floor, and that the wants of a particular constituency must be argued explicitly against other constituencies.

I think there should be a powerful representation of researchers in the organization. Not just a Science Officer, but almost a researcher congress, which is not merely called for “consultation”, but which pursues its own meetings, reports and conclusions. It should drive not only RAC, but organization-level emphasis, such as the kind of hardware configurations to purchase.

Lustre vs Librarians:

I am alarmed at the apparent lack of ambition at treating storage and archiving as parts of a whole. Everything I see about the curation/archiving side of things looks like handing your manuscript over to a librarian, who will bind it and put it onto a shelf. The problem is that archived data is useless: when ARC needs data, it should be on a live, mountable filesystem in native format. There seems to be little communication between FRDR-like efforts (which appear to be standalone portal-like systems) and what a researcher has most efficient and convenient access to on a cluster (mounted filesystems). This is not just a trivial difference of protocols, but what appears to me as a fundamental cultural outlook. I don’t know what the right approach is, but I think it’s a huge mistake to create separate organizations which will do their own divergent and largely non-interoperable things. At the same time, “one filesystem to rule them all” sounds like a disaster. I mention this as a problem not in any way to criticize the people working in these areas (including me!) but rather to point out that almost all ARC research is fundamentally tied to “data lifecycle”.