Service | Innovation | Value

Canadian Adventures in HPC Migration

iCAS 2017, September 12 2017
Luc Corbeil (SSC)
Alain St-Denis, Peter Silva (SSC)
Laurent Chardon (ECCC)

Agenda

- A bit of history
- A revamped HPC Solution
- Migration: opportunities and challenges

History

- Using HPC for weather forecasting since the 70's
- Previous contract with IBM
  - Initiated in 2003, Power/AIX, multiple upgrades
- Procurement launched in November 2014
  - Seeking for hosted supercomputing solution
    - Supercomputers
    - Pre-post processing clusters
    - Storage
    - Archiving
    - Network

HPC Renewal Contract Award

- New contract Awarded to IBM in May 2016
  - Installation phase
  - Initial solution, lease/hosting for 30-months
  - 2 performance upgrades at every 30 months
  - Optional 3rd performance upgrade

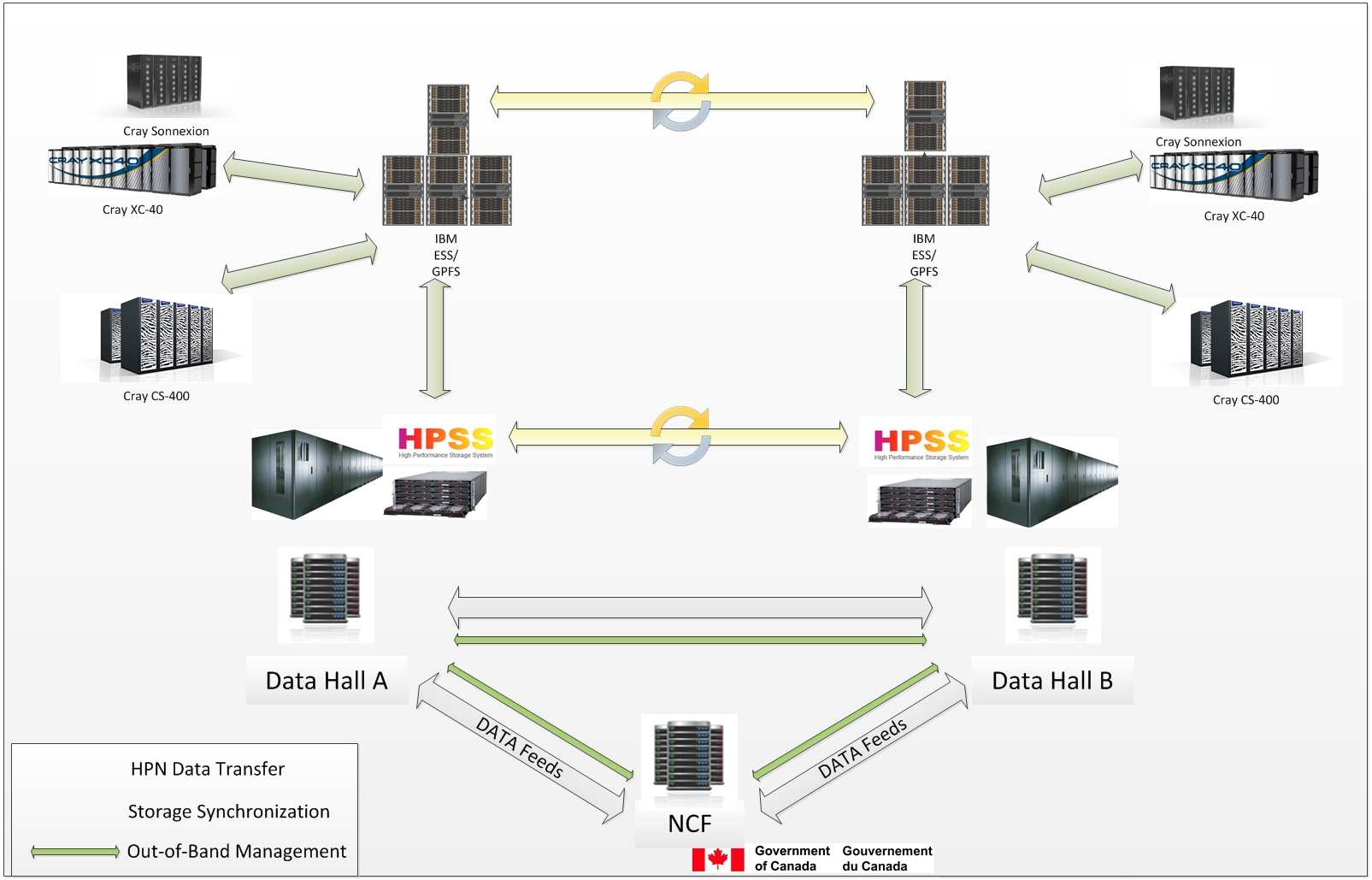
Initial Solution, Components (1)

- Two redundant data halls, each comprised of
  - Supercomputer
    - Cray XC-40, 856 computes nodes, Xeon E5-2695v4 18C 2.1GHz
    - 2.5 PB usable, Sonexion/Lustre storage
  - Pre/post processing (PPP) clusters
    - Cray CS-400, 158 compute nodes, Xeon E5-2699 v4 22C 2.2GHz
  - Site-store
    - IBM ESS GPFS storage, 18 PB

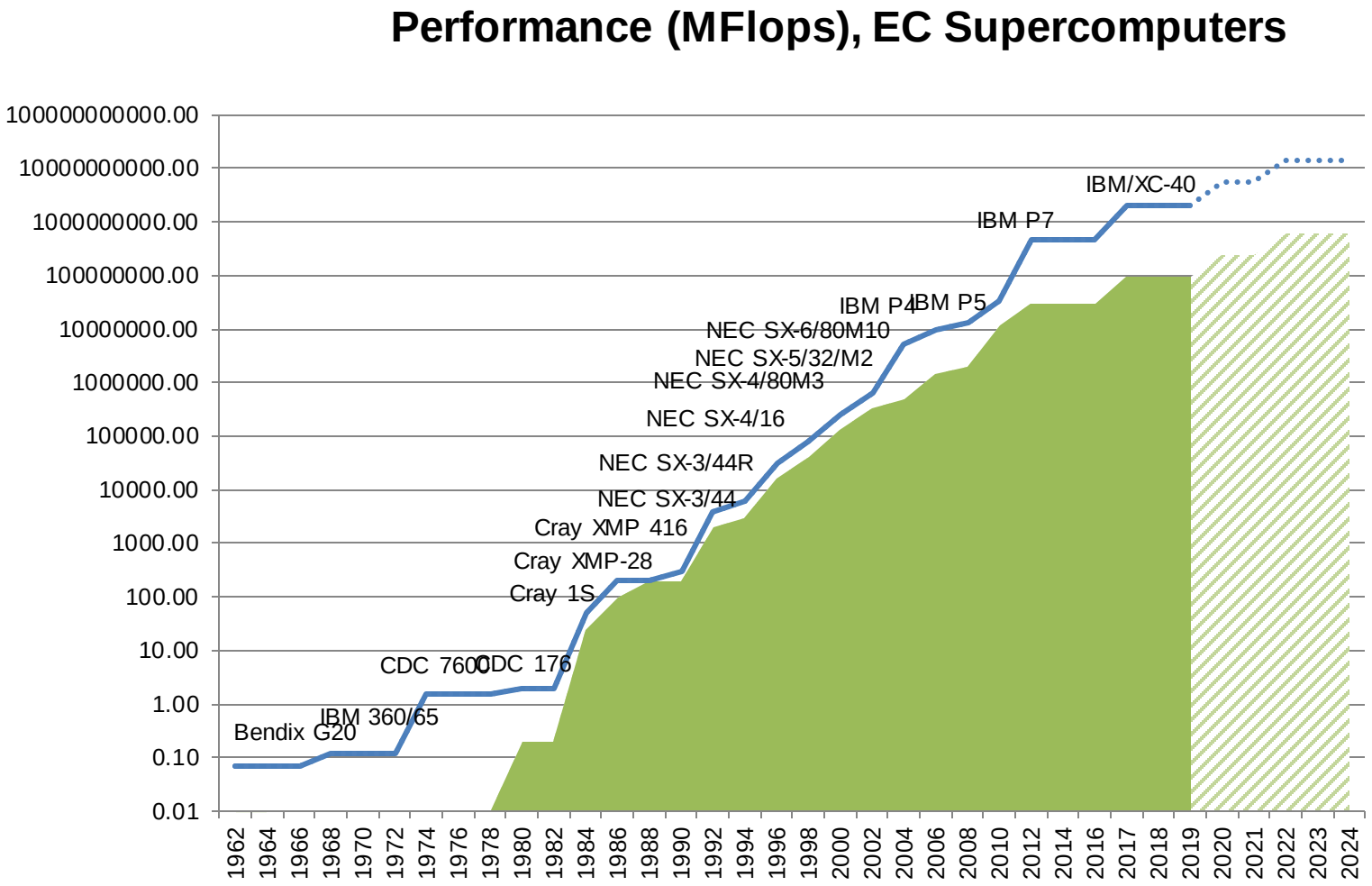
Initial Solution, Components (2)

- Shared accross halls
  - Homes
    - IBM ESS GPFS storage, full replication between halls
- HPSS for data archiving
  - 1.1 PB disk cache
  - 217 PB of tape storage (total of the two copies)

Resulting Architecture



The Latest Addition to a Long History



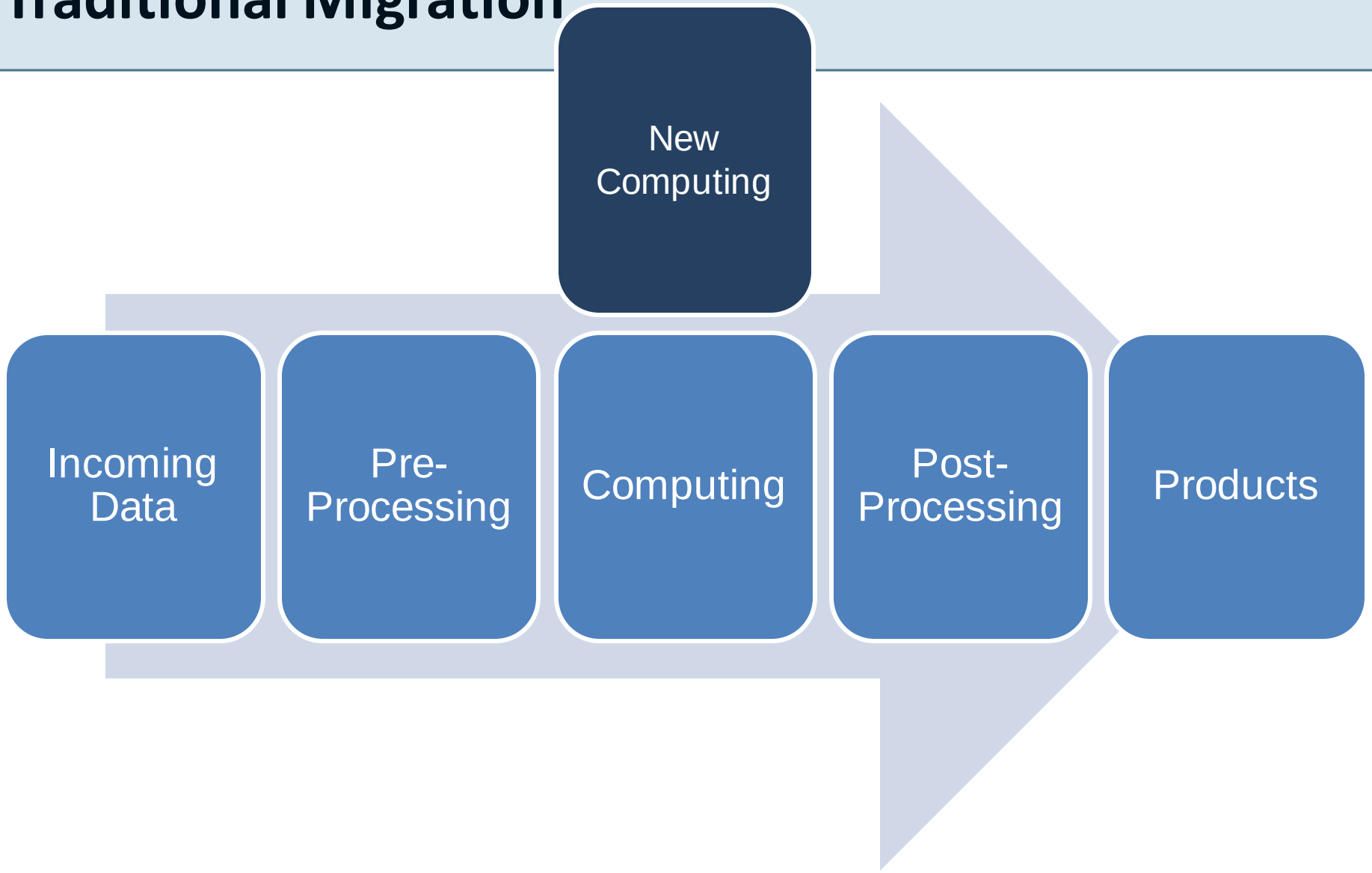
Migration: Key Aspects

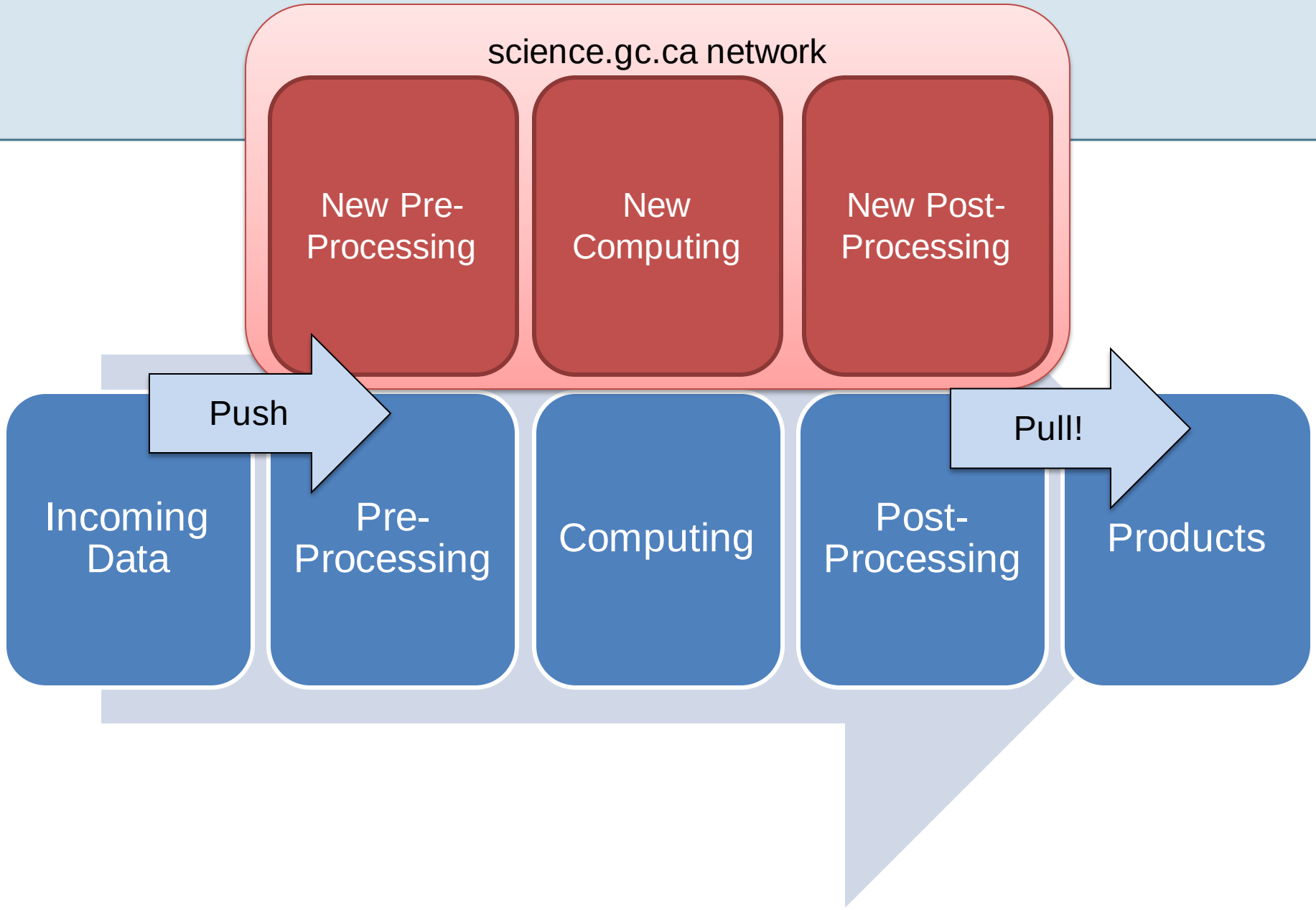
- Network & Environment
- Data movement
- Containers
- Timings

Network/Environment

- SSC's mandate covers all of Government of Canada scientific computing requirements
  - Bioinformatics, material sciences, etc.
- Opportunity
  - Create a networking zone dedicated to scientific computing
  - Install new HPC Solution for ECCC in this new environment
  - Foster more inter-departmental collaboration
- Challenge
  - Cannot create a cross-department network gateway

Traditional Migration




science.gc.ca network
New Pre-Processing
New Computing
New Post-Processing
Push
Pull!
Incoming Data
Pre-Processing
Computing
Post-Processing
Products

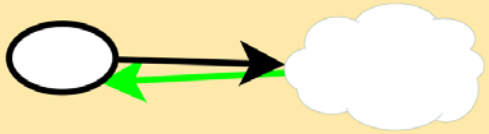
Data Movement: Sarracenia (1)

- MetPX-Sarracenia: Swiss Army Knife for data exchange
  - Based on AMQP (Advanced Message Queuing Protocol)
  - Used for data acquisition/dissemination, inter-hall mirroring
  - Network abstraction to the user
    - Data is advertised as ready at one end
    - Clients are advertising data they want to get
    - Sarracenia does the rest
  - http://metpx.sourceforge.net

Data Movement: Sarracenia (2)

Sarracenia Data Pumps http://metpx.sf.net

Data Sources: *sr_post*
Inject data once (or twice*)



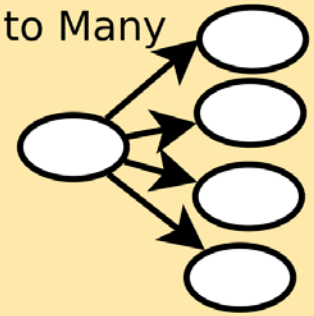
See when and where it went

Data Consumers: *sr_subscribe*

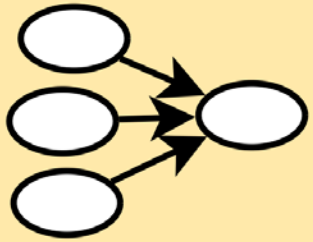
- faster notifications
- faster downloads
- no fixed IP needed
- self-serve

Sarracenia Benefits

Dissemination, replication: 1 to Many



Reliability through reduction: Many to 1



Simple performance through parallelism: multiple streams

Pump Network: Administrator Managed, Monitored, Optimized
For Administrators: Simpler than existing systems

File Movement Inside Solution (1)

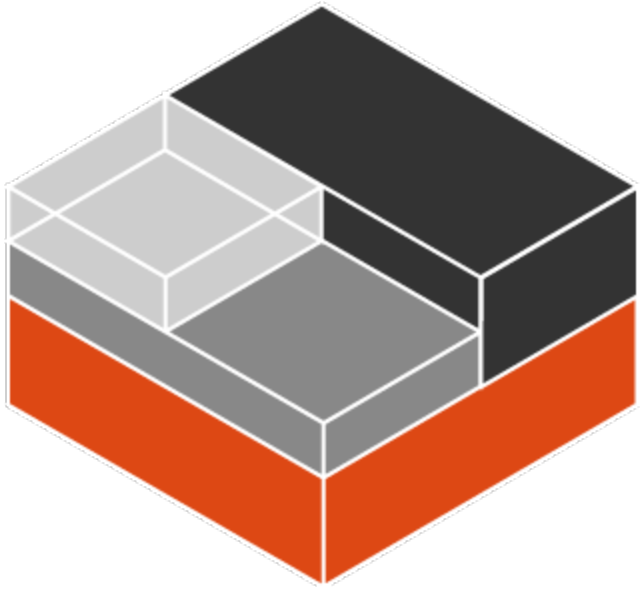
- Moving data is complex for users
  - « Feasible » is not equal to « optimal »
    - Many protocols to choose from (cp, scp, bbcp, rsync, ...)
    - Network link to use (IB native, IPoIB, 10Gb, etc...)
  - Consequence
    - On user: task takes longer than necessary
    - On overall infrastructure: server/network high load

File Movement Inside Solution: sscp

- Sscp
  - User just need to indicate from where to where
    - As long as ssh keys collaborate
  - Transfer mechanism selected at runtime
    - Considers data locality, file size, number of files, etc..

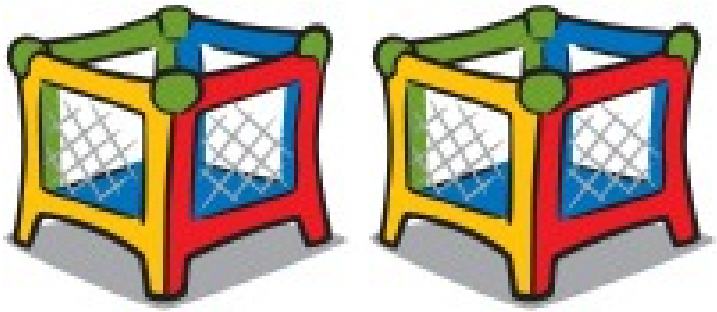
Completing a Migration Before it Starts

- Opportunity: do we need full solution on the floor to start migrating codes?
  - We can get a demo system
  - We can « call a friend » with a similar system
  - We can leverage containers
    - Destination OS/Compiler were known



Containers

- Ligthweight VMs, integrated with batch scheduler
  - User selectable Linux OS distro
  - Used to enforce resource limits (with cgroups)
    - Guarantees availability of requested resources at run time
  - Logical isolation
    - Users “see” only filesystems they have access to
  - No performance degradation, no emulation
  - Transparent to users



General Purpose Scientific Cluster

- Commodity cluster
  - Address computing requirements of science-based federal departments in Canada and allow them to collaborate
  - Using LXC containers to logically isolate various workloads in multi-tenant environment
  - Available prior to the installation of the PPPs

Impact on Migration

- Majority of migration work completed before PPP installation
  - Target OS/Environment was setup and deployed on GPSC
  - Through bind mounts, even filesystems names were set
  - Container replicated on PPP once installed
- Final validation of applications PPP became trivial
- Future OS migrations can go on a per-app basis
  - No more D-Day where everything has to move at once

Containers: Future Work

Potential in Singularity

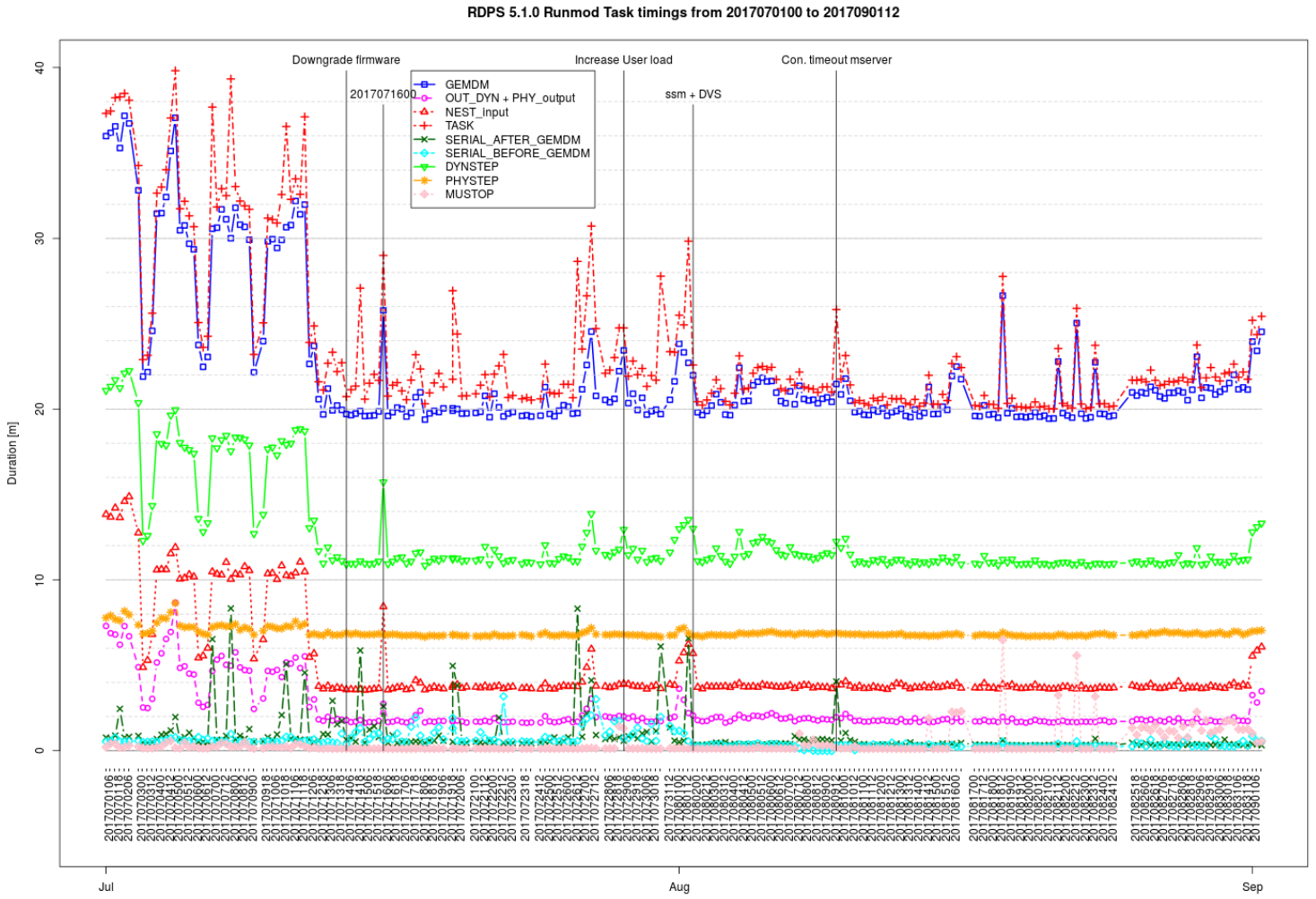
- Relieve sysadmins from Linux distro image management
  - Singularity puts image management in the hands of the users, BYOI
- Simplify local extensions to the batch scheduler system profile
- Remove the SR-IOV layer (virtual split of PCI network devices)
  - Singularity uses the host's network namespace
- Security: runs in user space.
  - Image management requires privileges: done elsewhere

Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: Scientific containers for mobility of compute. PLoS ONE 12(5): e0177459. https://doi.org/10.1371/journal.pone.0177459

Timings

- Users love consistent application performance
- Disturbance can come from many sources
- Crucial for Ops to have predictive timings
  - Must meet SLA for product delivery timeliness

Hunting for Consistent Performance



*Time series of RDPS performance, André Plante, ECCC

Conclusion

- A new HPC Solution has been deployed
- Significant improvements, not only in performance
  - Architecture: network, environment, redundancy
  - Data movement
  - Customization of the platform
- Next steps
  - Accelerators
  - Visualization
  - Improved Containers

Questions?