OPENFABRICS ALLIANCE

Scalable Subnet Administration

Hal Rosenstock
Mellanox Technologies

The Problem

n^2 SA load

- SA queried for every connection
- Communication between all nodes creates an n^2 load on the SA
  - In InfiniBand architecture (IBA), SA is a centralized entity
- Other n^2 scalability issues
  - Name to address (DNS)
    - Mainly solved by a hosts file
  - IP address translation
    - Relies on ARPs

A Truly Novel Solution…

OPENFABRICS ALLIANCE

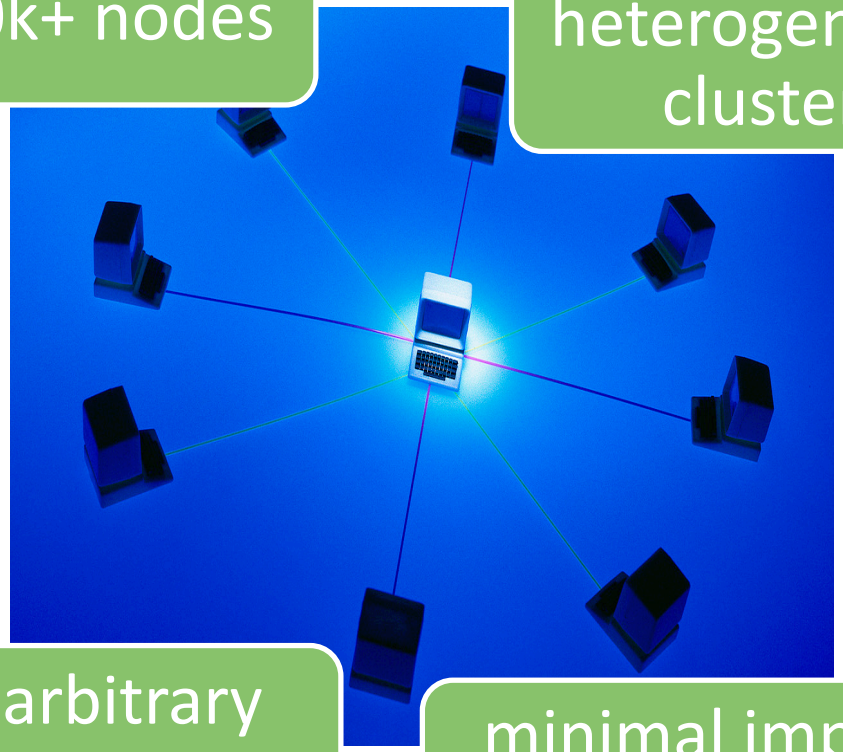
Scalable SA (SSA)

- Extends the SA implementation
  - Improved opportunities for solutions
  - Open source
- Focused on scalability *and* reliability
  - Fault occurrence is likely
- Dependent on SM
- Turns a centralized problem into a distributed one

Goals

40k+ nodes

heterogeneous clusters

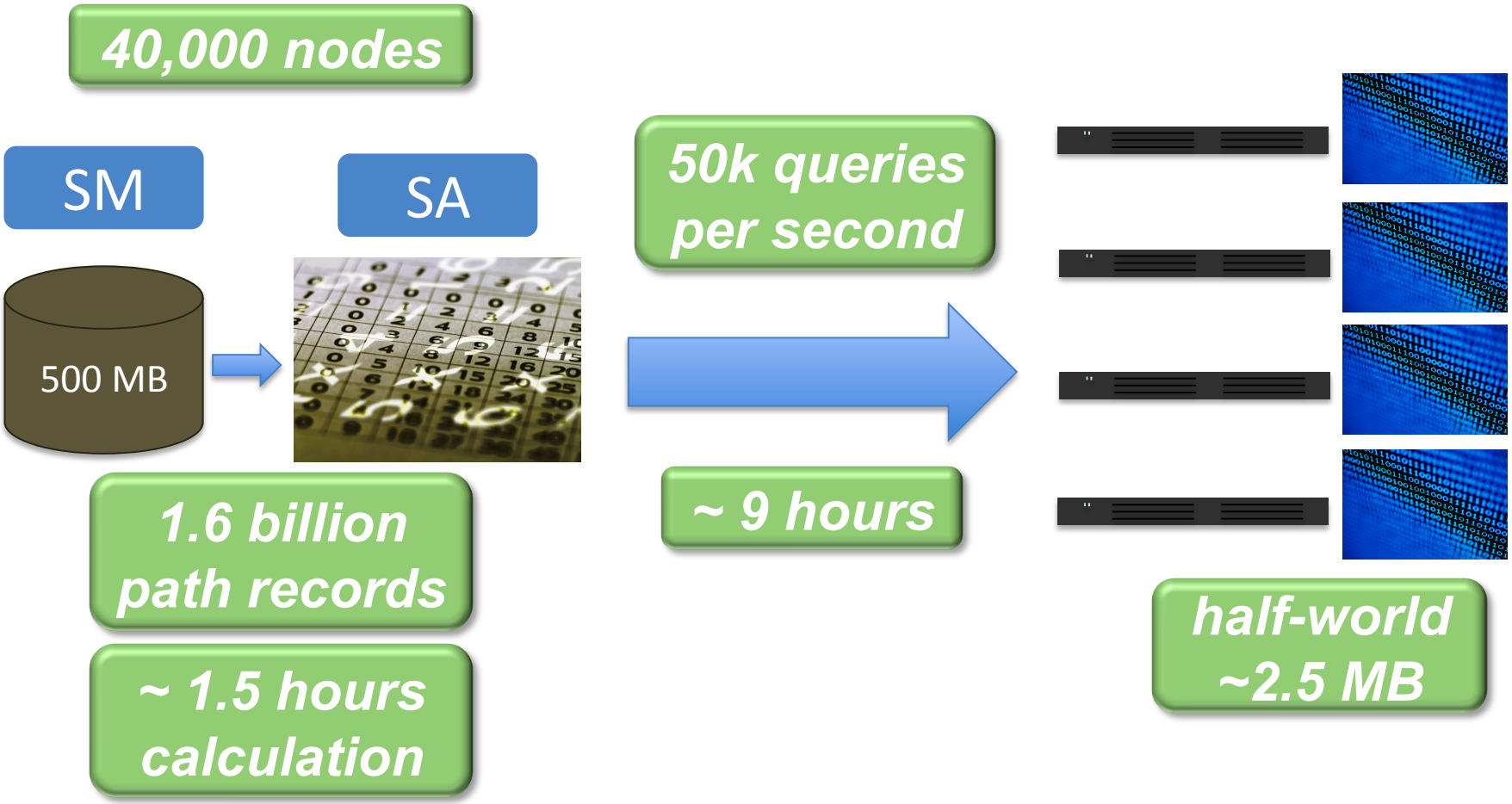


arbitrary topologies

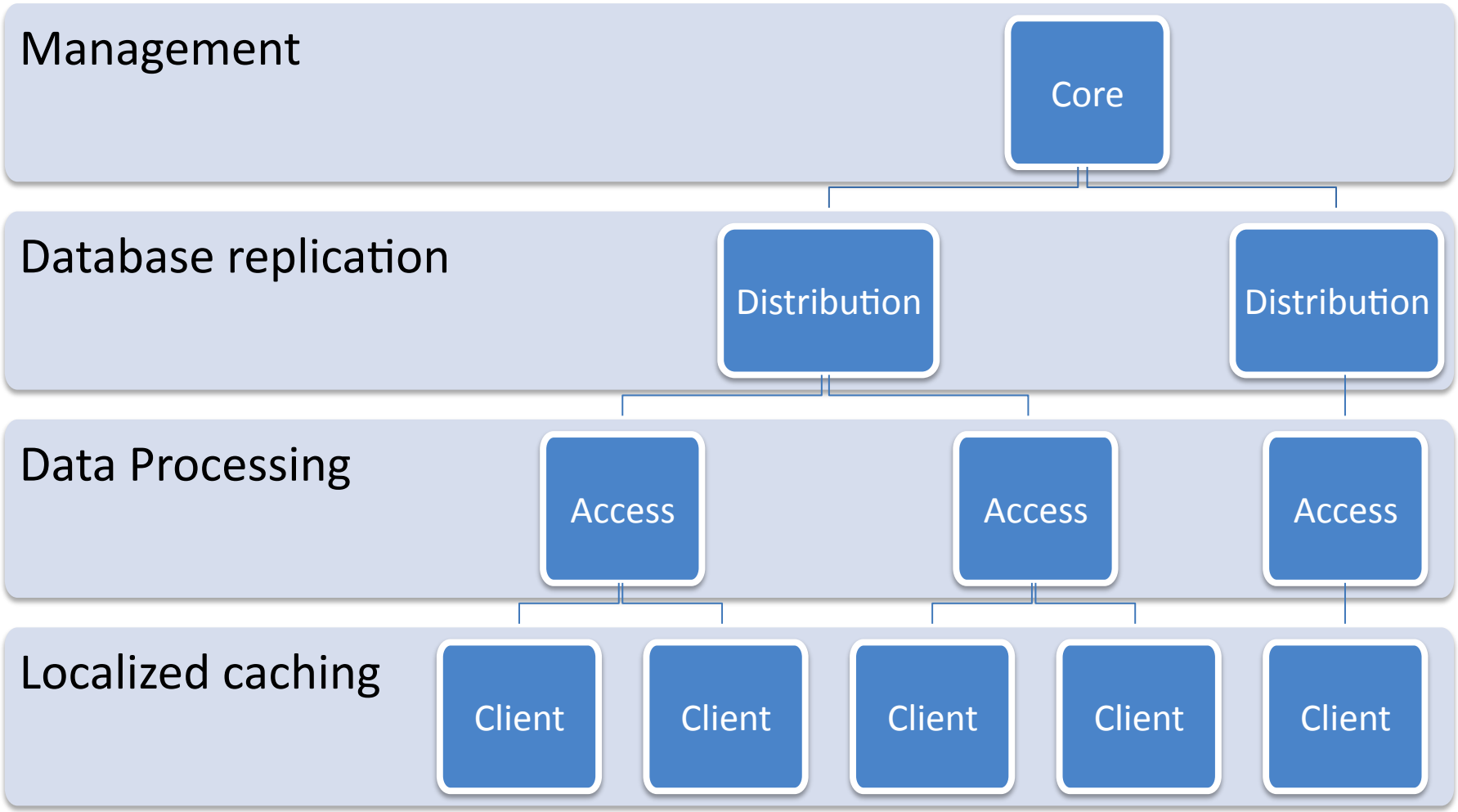
minimal impact to compute nodes

- Support distributed processing
- Ensure consistency of cached data
- Avoid large multicast domains
  - Do not rely on IPoIB
- Work with existing RDMA CM apps

Analysis


40,000 nodes
SM
SA
500 MB
1.6 billion path records
~ 1.5 hours calculation
50k queries per second
~ 9 hours
half-world ~2.5 MB

SSA Architecture



Distribution Tree

- Number of management nodes needed is dependent on subnet size and node capability (CPU speed, memory)
  - Combined nodes
    - Core and access
    - Distribution and access
- Fanouts in distribution tree for 40K compute nodes
  - 10 distribution per core
  - 20 access per distribution
  - 200 consumer per access

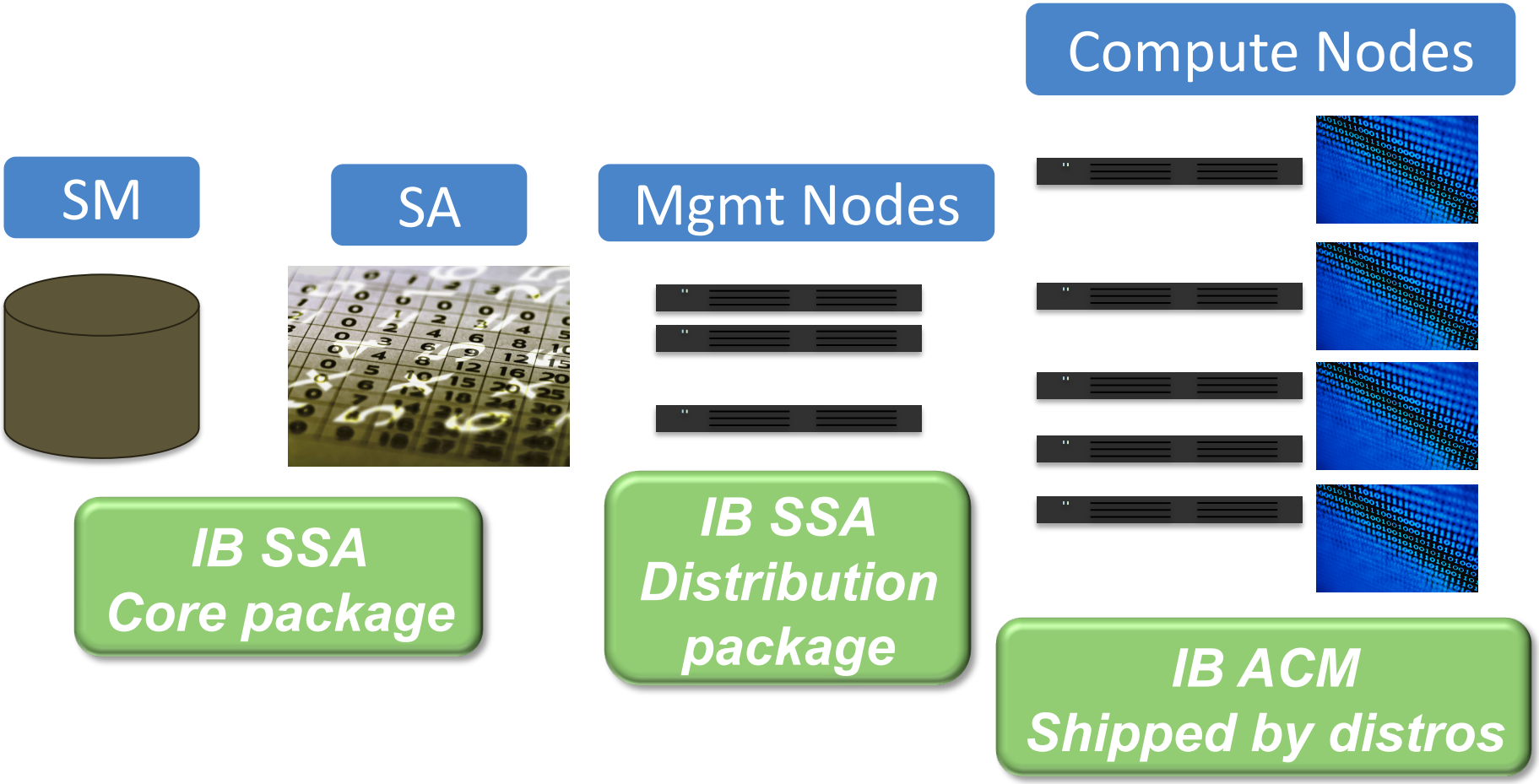
System Requirements

- AF_IB capable kernel
  - 3.11 and beyond
- librdmacm with AF_IB and keepalive support
  - Beyond 1.0.18 release
- libibverbs
- libibumad
  - Beyond 1.3.9 release
- OpenSM
  - 3.3.17 release or beyond

OpenMPI

- RDMA CM AF_IB connector contributed to master branch recently
  - Thanks to Vasily Filipov @ Mellanox ☺
  - Need to work out release details
    - Not in 1.7 or 1.6 releases

Deployment


Compute Nodes
SM
SA
Mgmt Nodes
IB SSA
Core package
IB SSA
Distribution
package
IB ACM
Shipped by distros

Project Team

- Hal Rosenstock (Mellanox)
- Sean Hefty (Intel)
- Ira Weiny (Intel)
- Susan Coulter (LANL)
- Ilya Nelkenbaum (Mellanox)
- Sasha Kotchubievsky (Mellanox)
- Lenny Verkhovsky (Mellanox)
- Eitan Zahavi (Mellanox)
- Vladimir Koushnir (Mellanox)

Initial Release

- Path Record Support
- Limitations (Not Part of Initial Release)
  - QoS routing and policy
  - Virtualization (alias GUIDs)
- Preview – June
- Release - December

Summary

- A scalable, distributed SA
- Works with existing apps with minor modification
- Fault tolerant

- Please contact us if interested in deploying this

OPENFABRICS
ALLIANCE