Mach11

Customer Experiences

Andreas Weininger
Andreas.Weininger@de.ibm.com

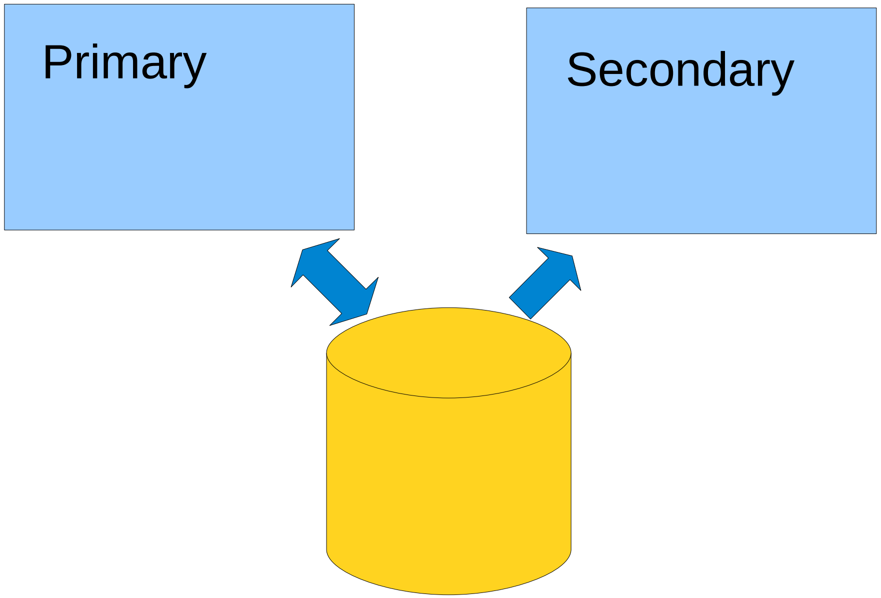
Infobahn 2009

Agenda

- Examples of Use Scenarios for Shared Disk Secondaries
  - SDS for HA
  - SDS for Scalability
  - SDS for Workload Isolation
- Case Study: A Mach 11 Cluster at a German Bank
  - Problem
  - Design of the Mach 11 Cluster
  - Performance
  - High Availability
  - Migration
  - Lessons Learned

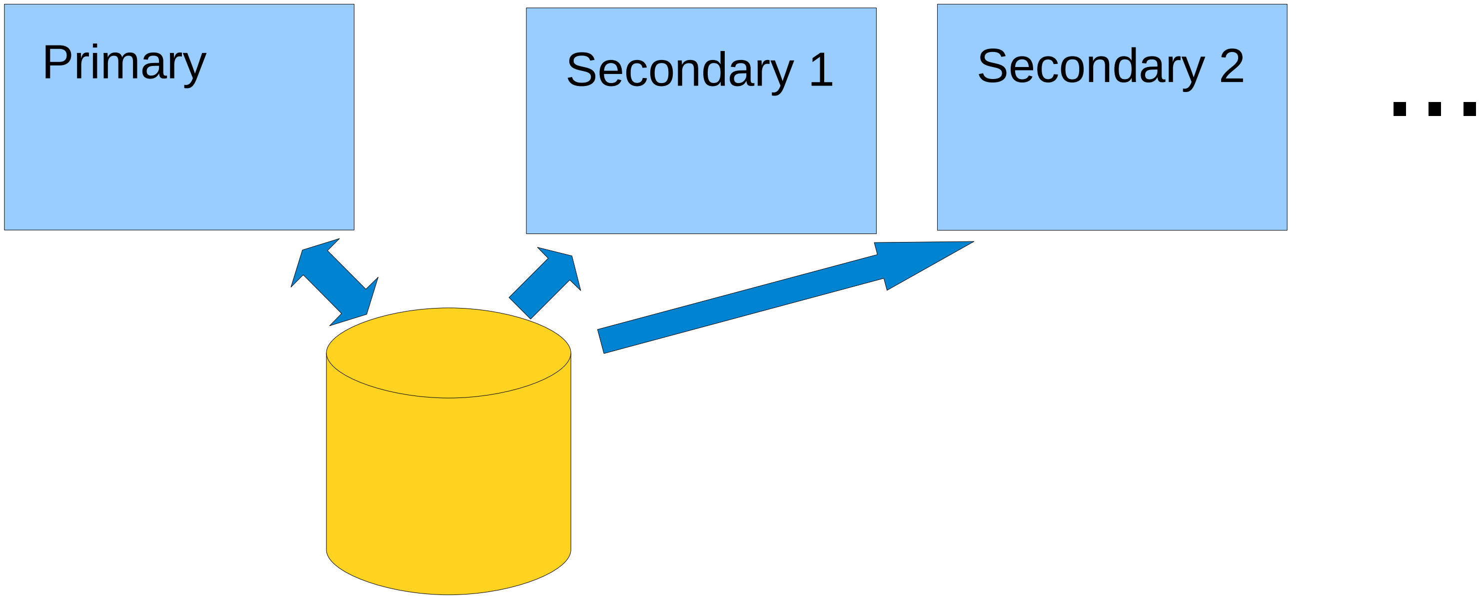
Examples of Scenarios for Shared Disk Secondaries

Shared Disk Secondaries for High Availability



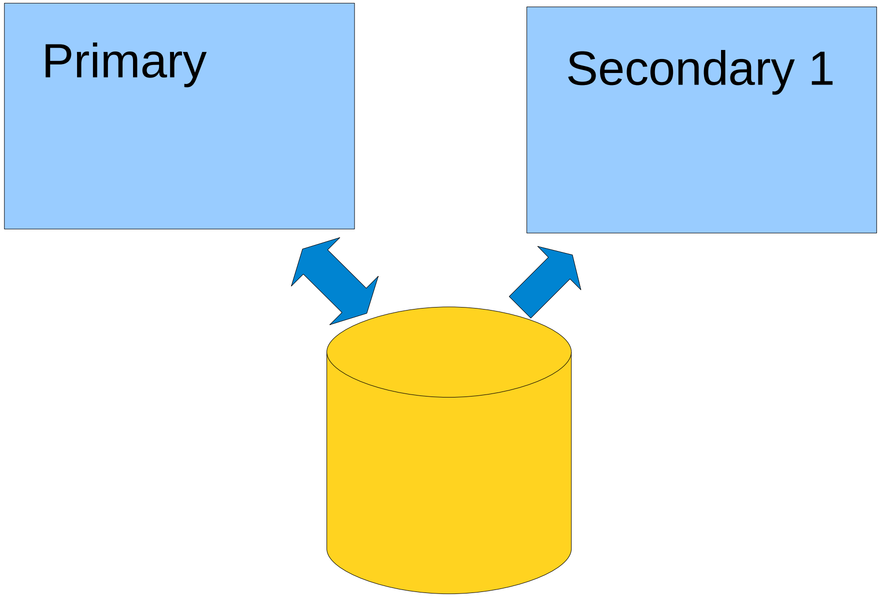
- Advantages
  - Very easy to set up
  - Very fast failover

Shared Disk Secondaries for Scalability



- Advantages
  - Easy online extension of cluster
  - Number of Secondaries according to workload requirements

Shared Disk Secondaries for Workload Isolation



- Advantages
  - Workload with unknown characteristics can be separated on other system e.g. ad hoc queries
  - Minimization of performance impact on primary
  - But still same view of data as primary

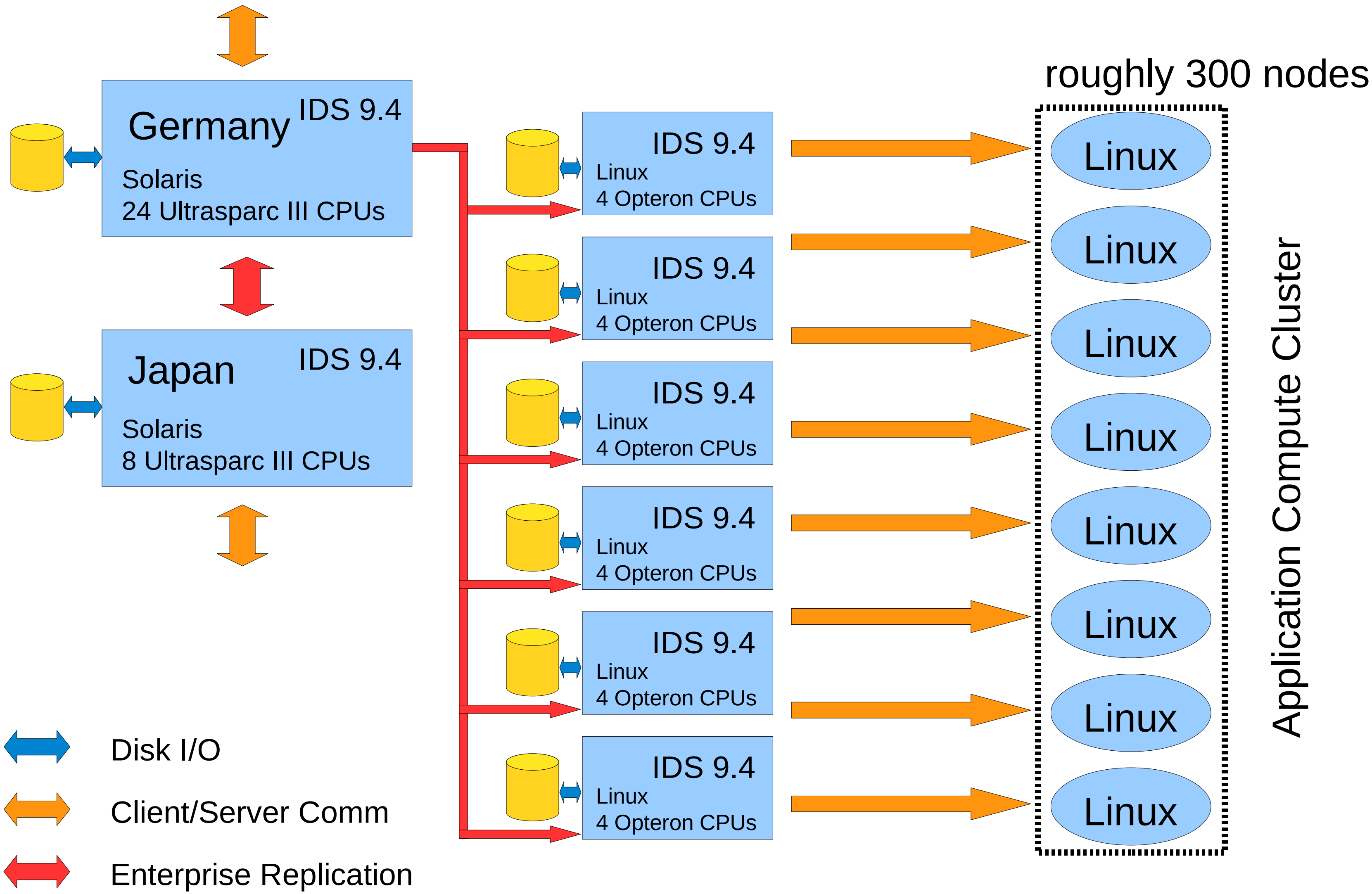
Case Study:
A Mach11 Cluster at a German Bank

Problem

Application

- Analysis of securities
- Very computation intensive
- Each user gets 5 to 20 nodes on a Linux compute cluster
- Application is fully parallelized to use all the nodes
- Each node connects to database to get information on the securities (access is read-only)
- New information about securities continuously inserted
- Application is very business critical; therefore high availability requirements

Old Architecture



Evaluation of Existing Architecture

Pros:

- good availability
- fast disaster recovery
- good scalability
- sufficient performance

Cons:

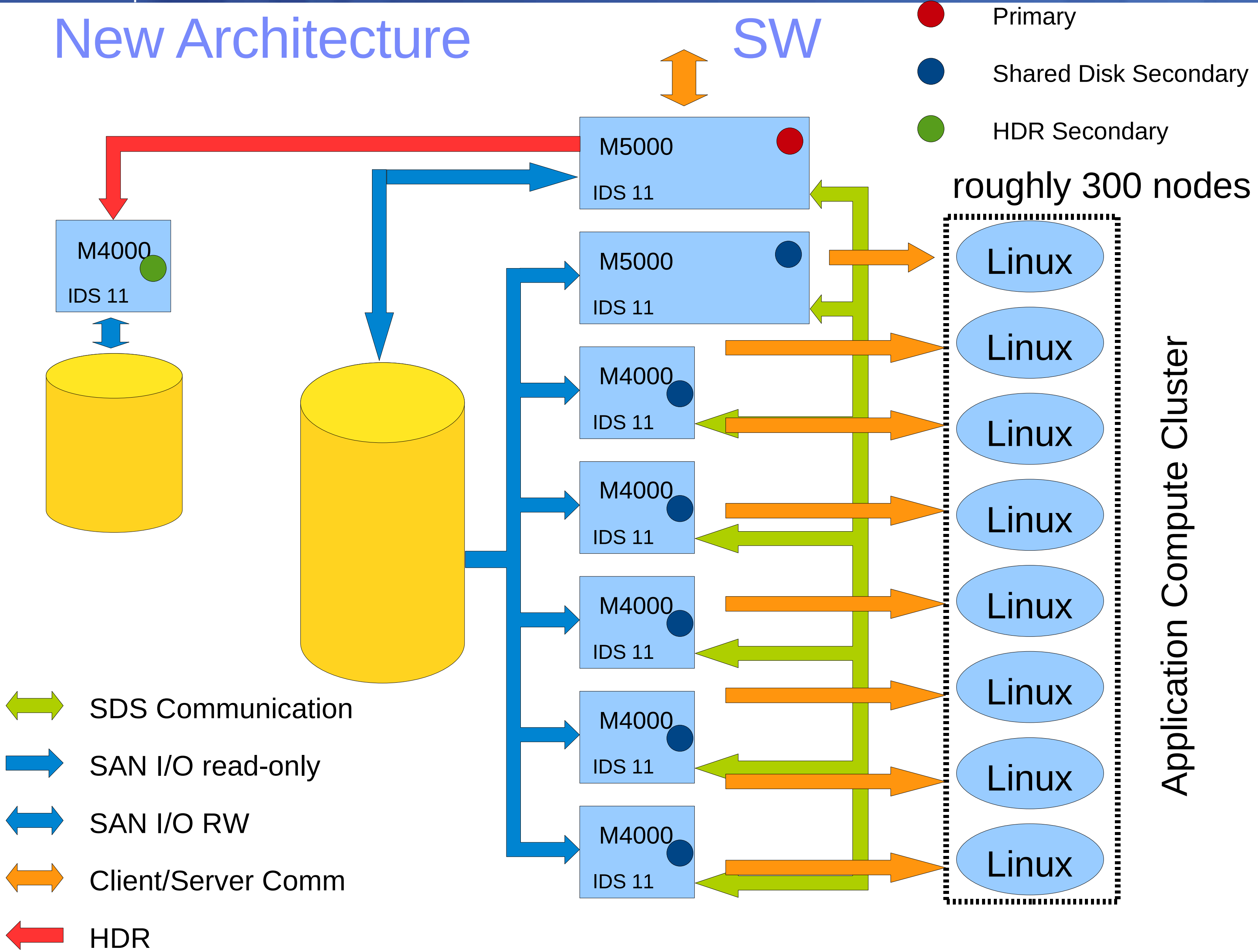
- enterprise replication requires significant administration efforts
- costs

Design Mach 11 Cluster

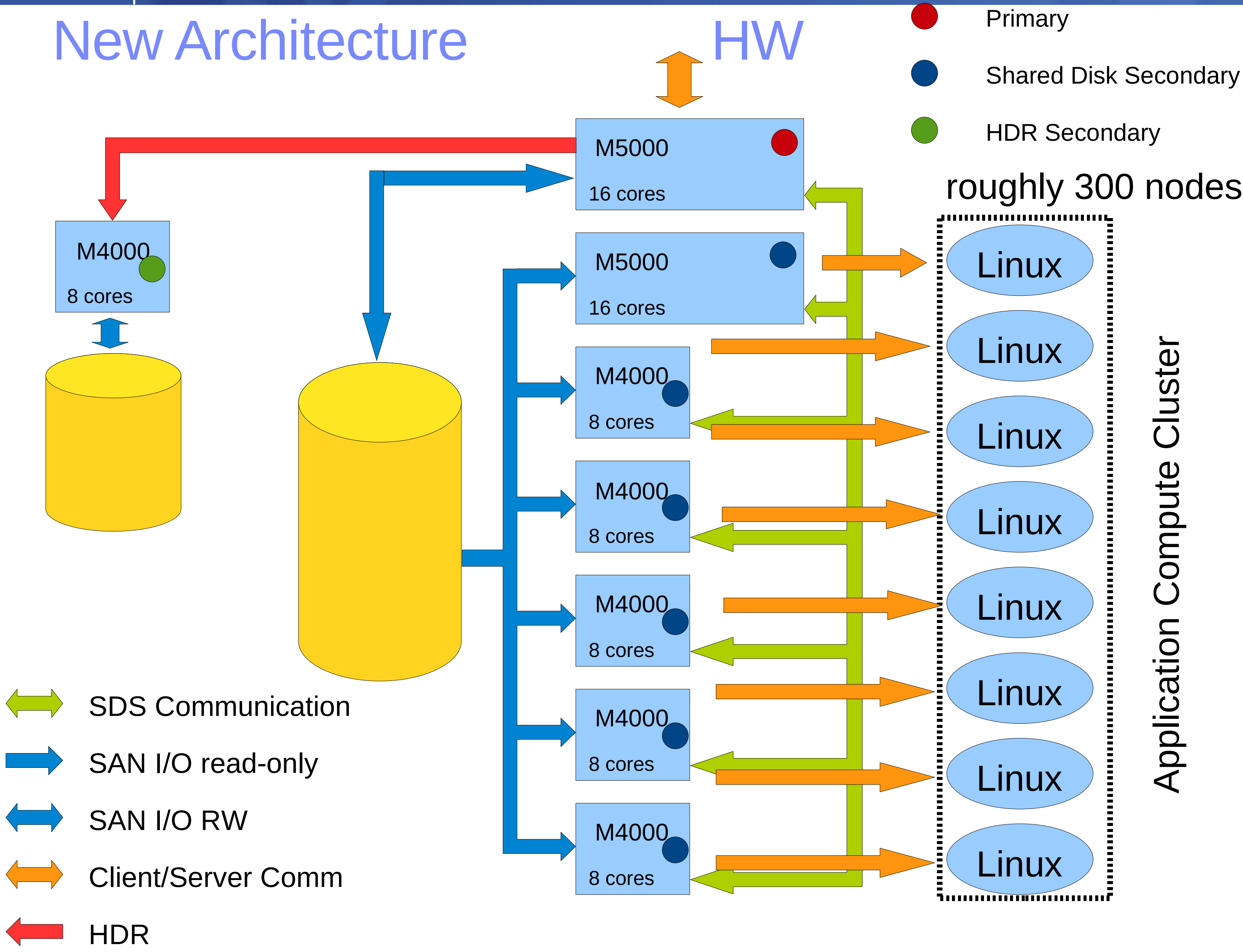
Requirements and Goals

- Scalability
- High availability
- Fast disaster recovery
- Minimize changes to application
- Smooth transition from old system
- No Linux for Database Servers
- Improvement of TCO
  - Reduce Maintenance Costs
  - Minimize Software License Costs (especially for 3rd Party Software)

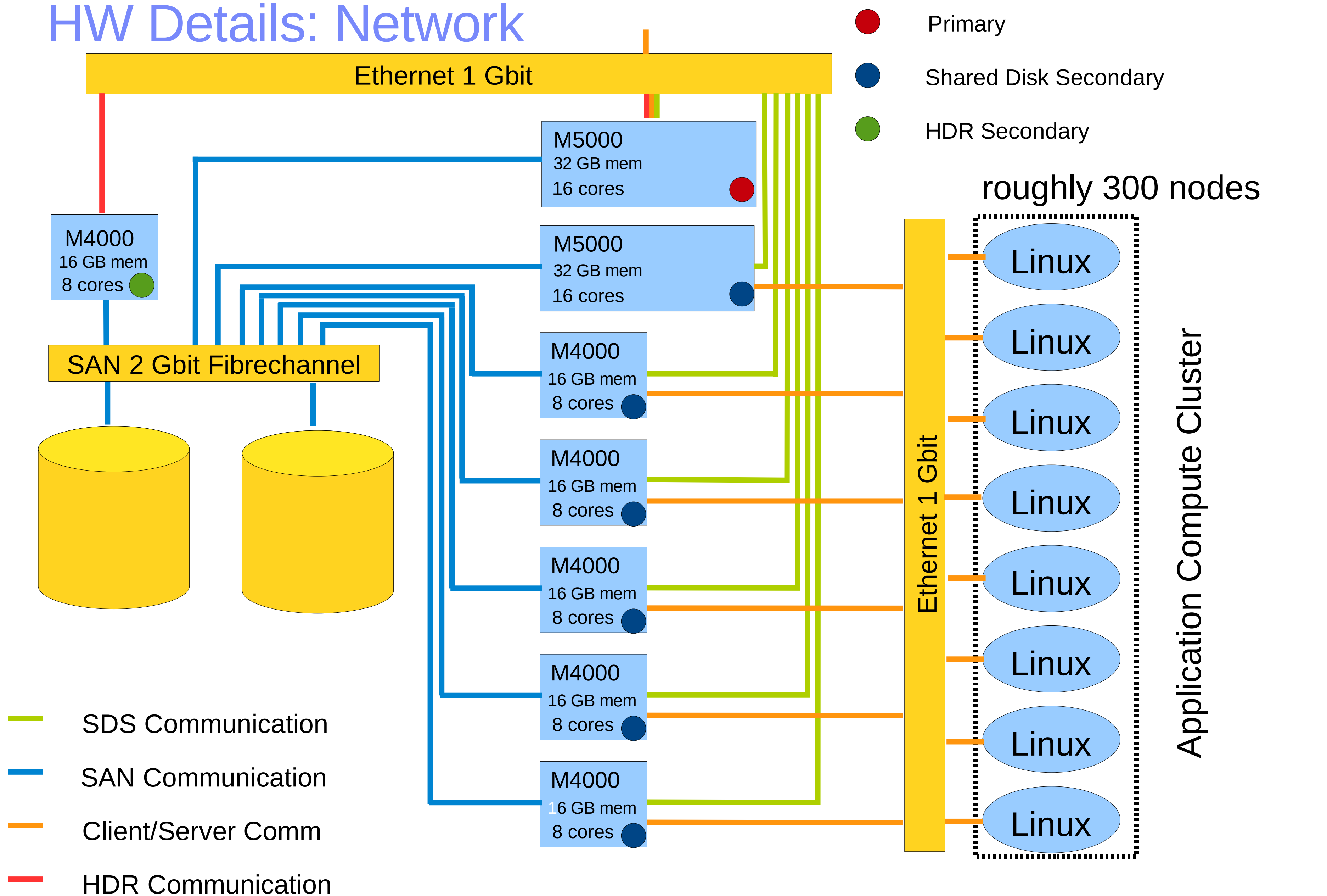
New Architecture



New Architecture HW



HW Details: Network

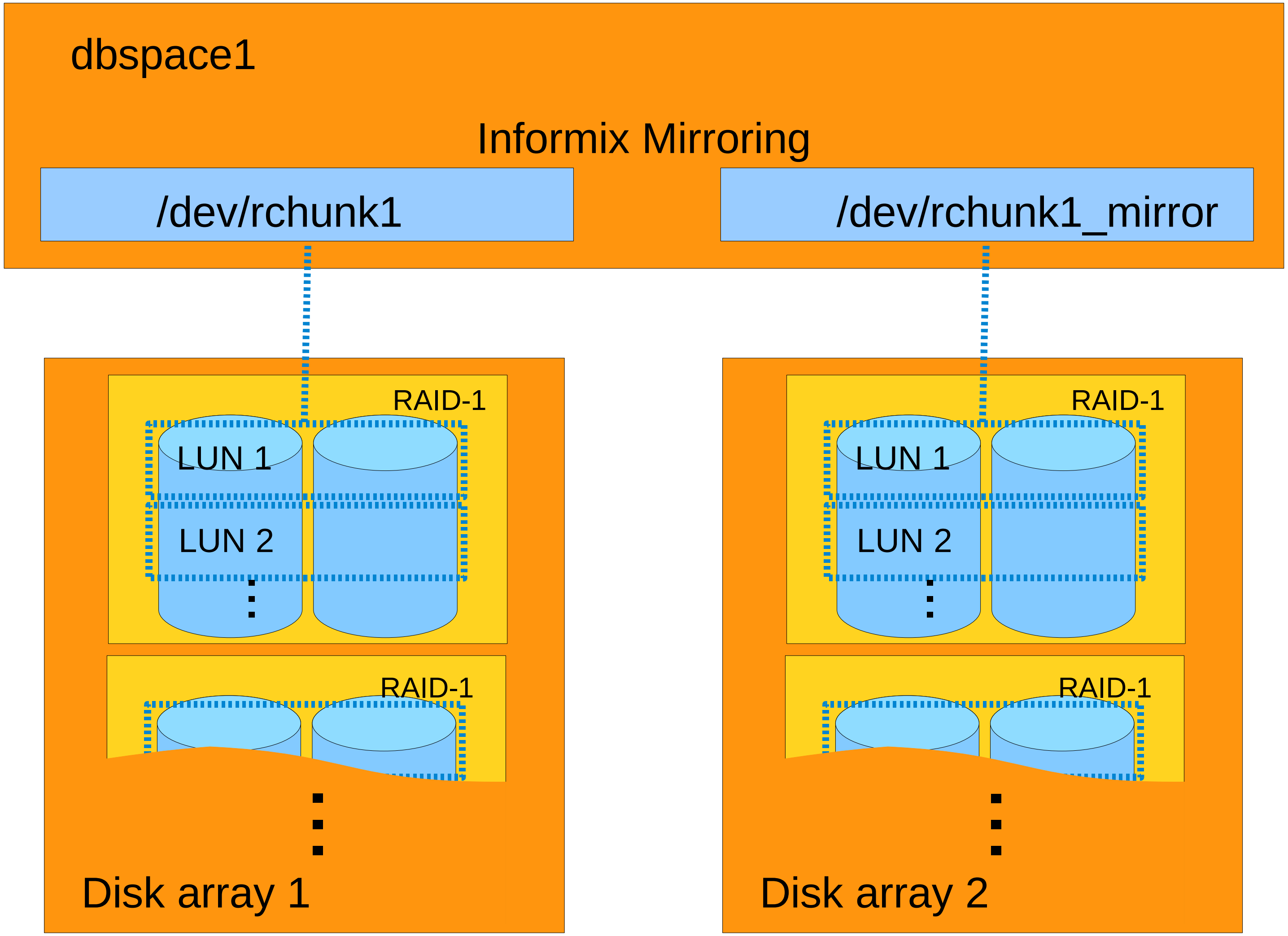


Redundancy at the Disk Level

- Primary and Shared Disk Secondaries share one set of logical dbspaces D1
- Local dbspaces of secondaries also located on SAN
- HDR Secondary has second copy of these dbspaces D2
- For all dbspaces in D1 and D2 Informix mirroring is used i.e. for each chunk c_i there is a mirror chunk cm_i
- The chunks c_i and cm_i are mapped to LUNs in two physically different disk arrays
- RAID-1 is used for each LUN

Mapping of DBSpaces to Disks

visible on all nodes

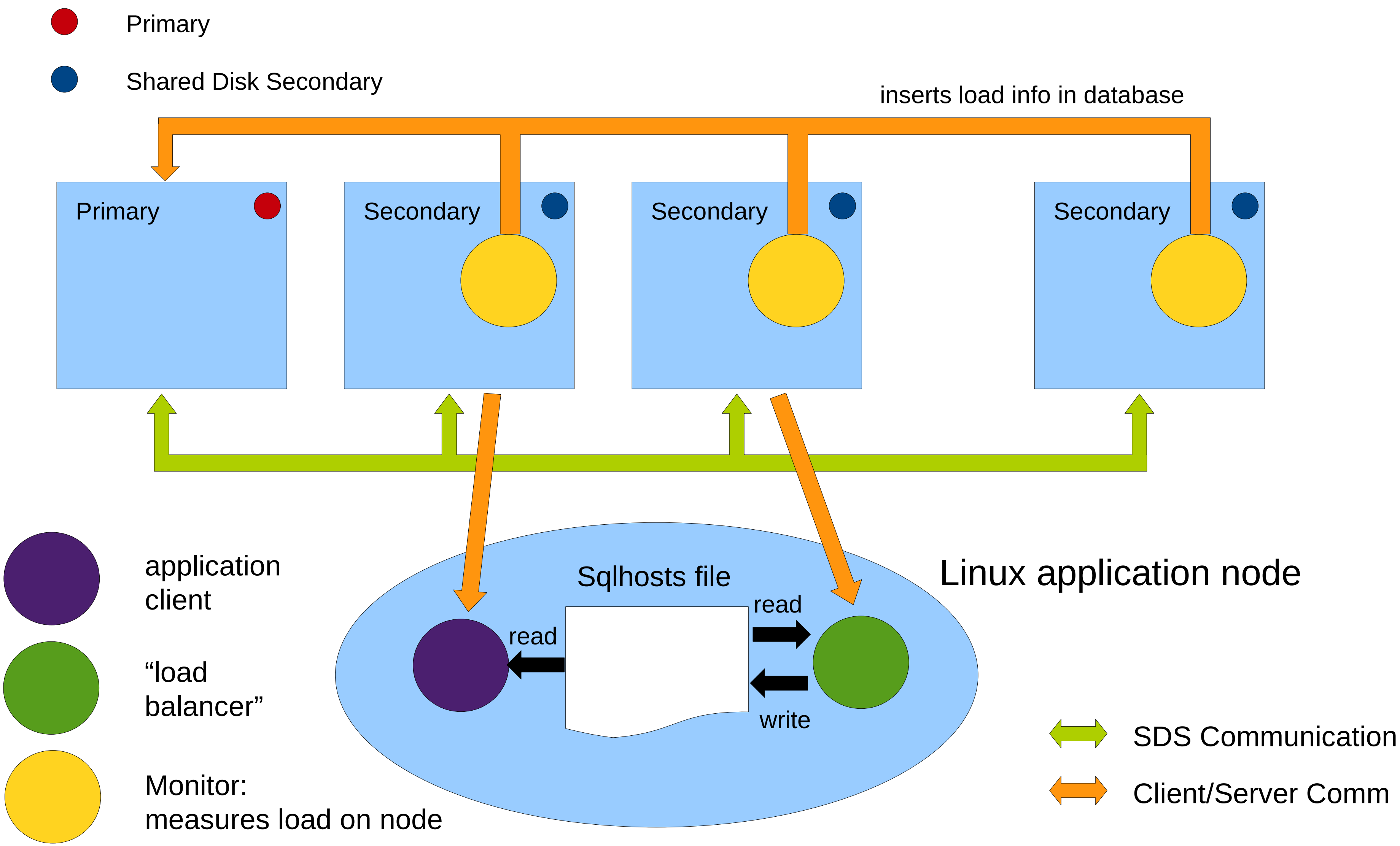


raw devices

Why was the mapping of the dbspaces done this way?

- High degree of availability (see discussion on later slides)
- Good performance
- Raw devices instead of shared file system:
  - Cost of shared file system avoided
  - Performance
  - Stability (avoid additional SW layers)
  - Ease of use
- IDS Mirroring instead of LVM mirroring
  - Cost of logical volume manager avoided

Communication with the Compute Cluster



Performance

Avoiding Bottlenecks

- Shared Disk subsystem
- SAN
- Number of Cores
- Memory
- Ethernet

Shared Disk subsystem / SAN

- Shared Disk subsystem has to provide sufficient I/O bandwidth and number of I/Os for primary and all shared disk secondaries (do not size by disk capacity)
- Example:
  - Primary:
    - 400 MB/s bandwidth
    - 2000 IO/s
  - Each secondary (6 secondaries):
    - 200 MB/s bandwidth
    - 1500 IO/s
  - Requirements for shared disk subsystem:
    - > 1600 MB/s bandwidth (400+6*200)
    - > 11000 IO/s (2000+6*1500)

Scalability: Adding CPUs and Nodes

- Read-Write Load:
  - Options for Scaling: additional CPUs per node
  - Distributed Writes: If IUD operation is very compute intensive
- Read-Only Load:
  - Options for Scaling: additional CPUs per node
    - e.g. M5000 with 16 cores instead of 8 cores
  - Options for Scaling: additional nodes
    - Size may vary, but slowest secondary may determine maximum throughput of primary
  - Options for Scaling: Read-Only Clients also on HDR secondary

Memory and Ethernet

- Amount of memory:
  - Is working set of primary and secondaries similar?
  - Sizing of memory based on old system
- Ethernet:
  - GBit Ethernet especially for HDR
  - Not much bandwidth requirements for SDS

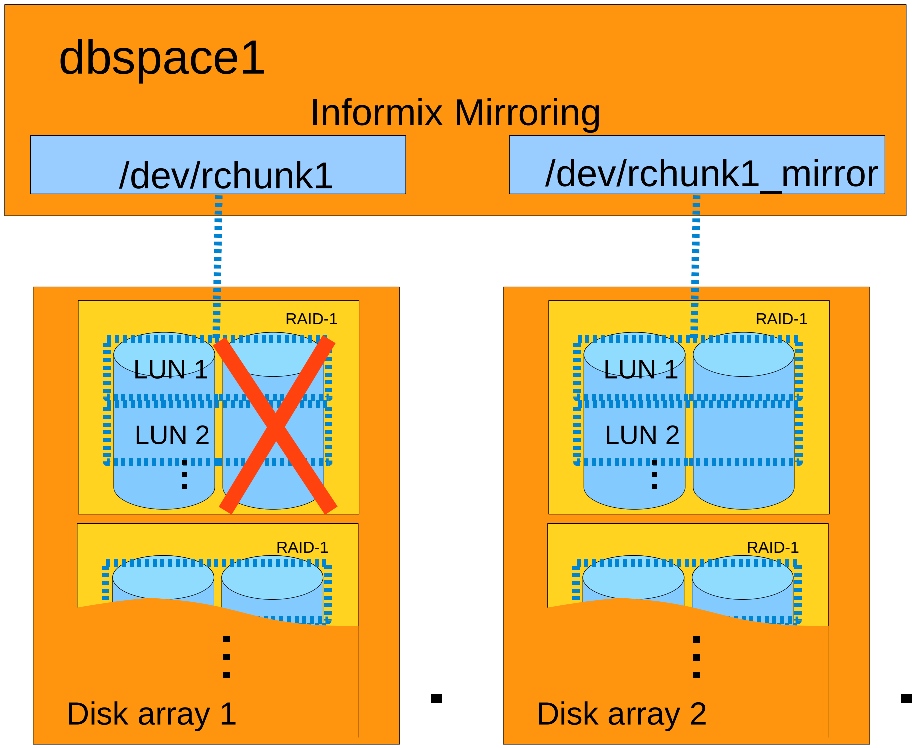
High Availability

Different Availability Scenarios

- Loss of physical disk
- Loss of disk array
- Loss of network (SAN or Ethernet both not discussed)
- Loss of primary
- Loss of HDR secondary
- Loss of SD secondary
- Loss of whole data center
- Corruption of shared disk
- Scheduled maintenance

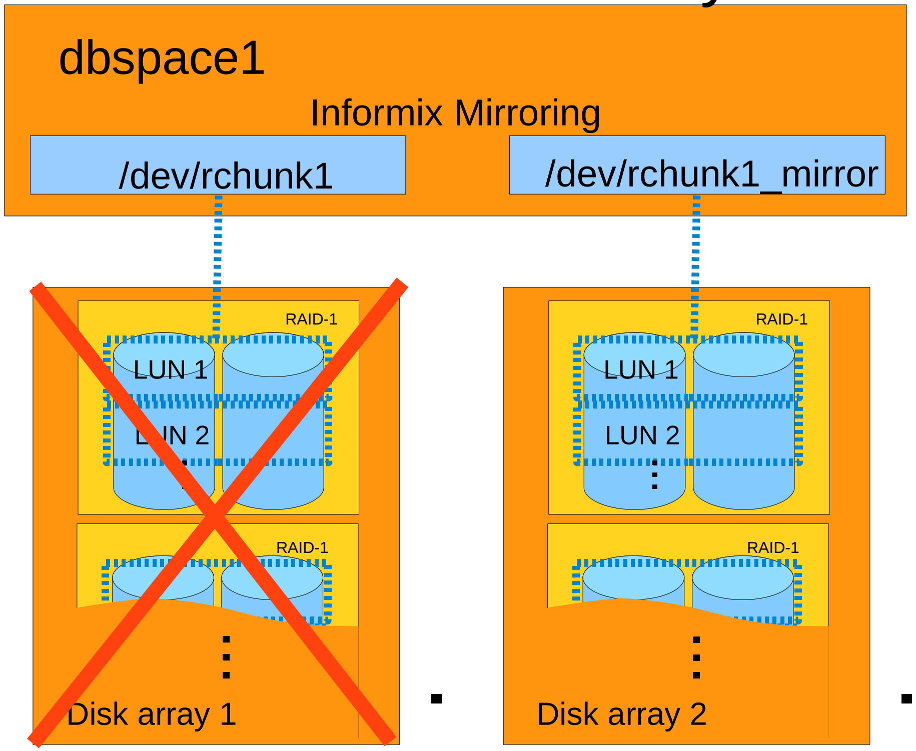
Loss of a physical disk

- Captured locally in disk array
- Replacement and resilvering
- Transparent to all node and IDS



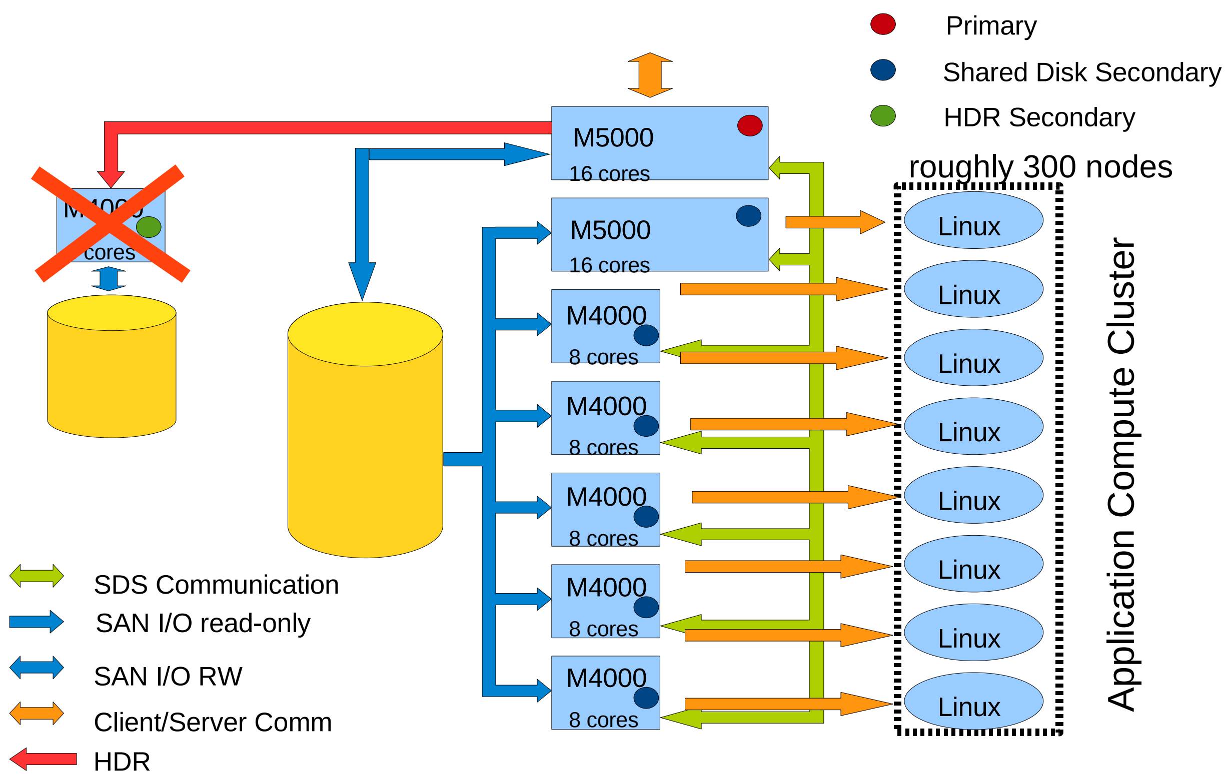
Loss of a disk array

- Captured by Informix mirroring
- Chunk and mirror chunk on different disk arrays
- All mirrors are lost in case of disk array failure



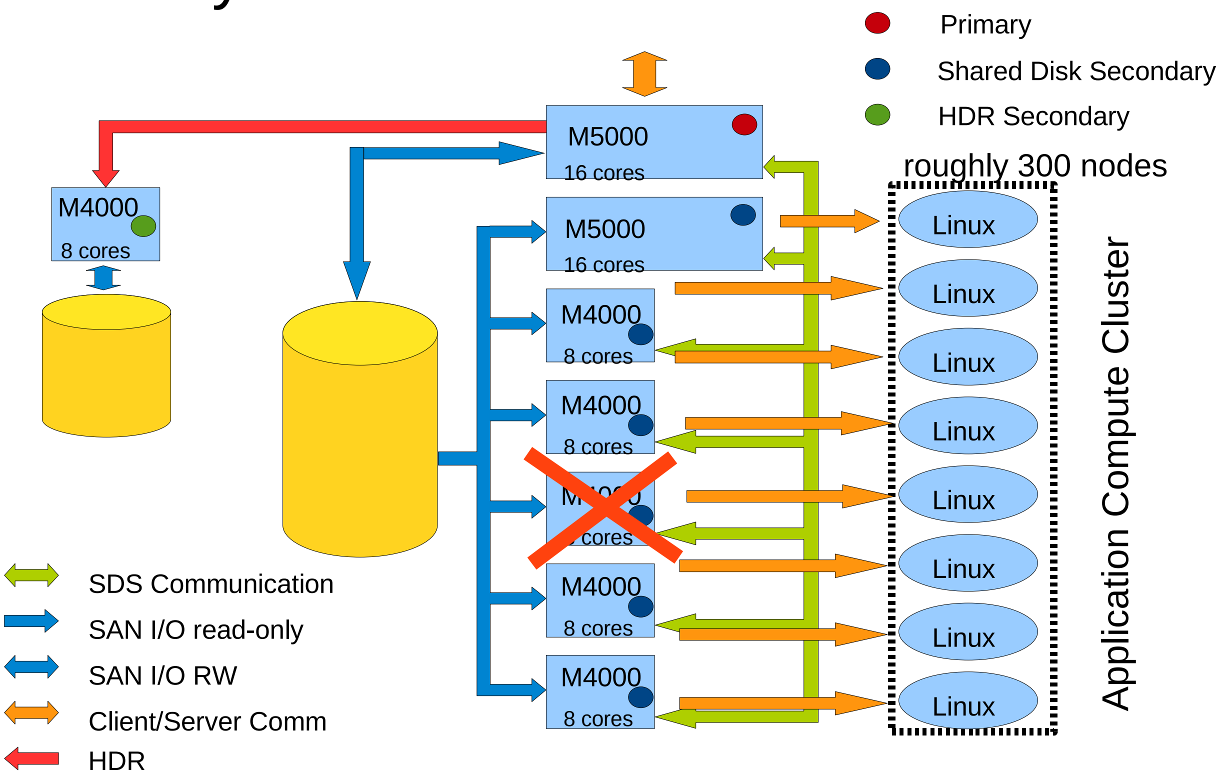
Loss of HDR Secondary

- No direct impact on clients since they aren't connected to the HDR secondary
- Only potential availability reduced



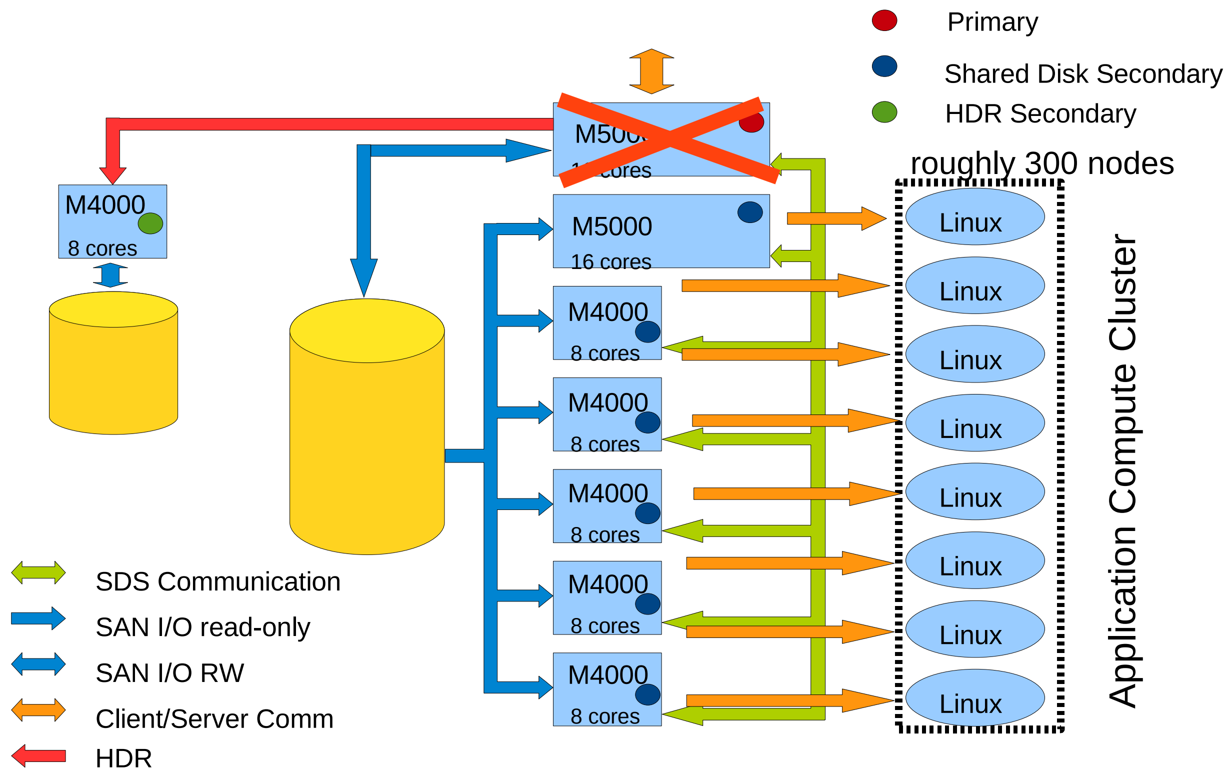
Loss of Shared Disk Secondary

- No immediate impact on remaining servers
- Clients of failed node have to reconnect to other SDS node
- Node no longer used by load balance
- Performance impact



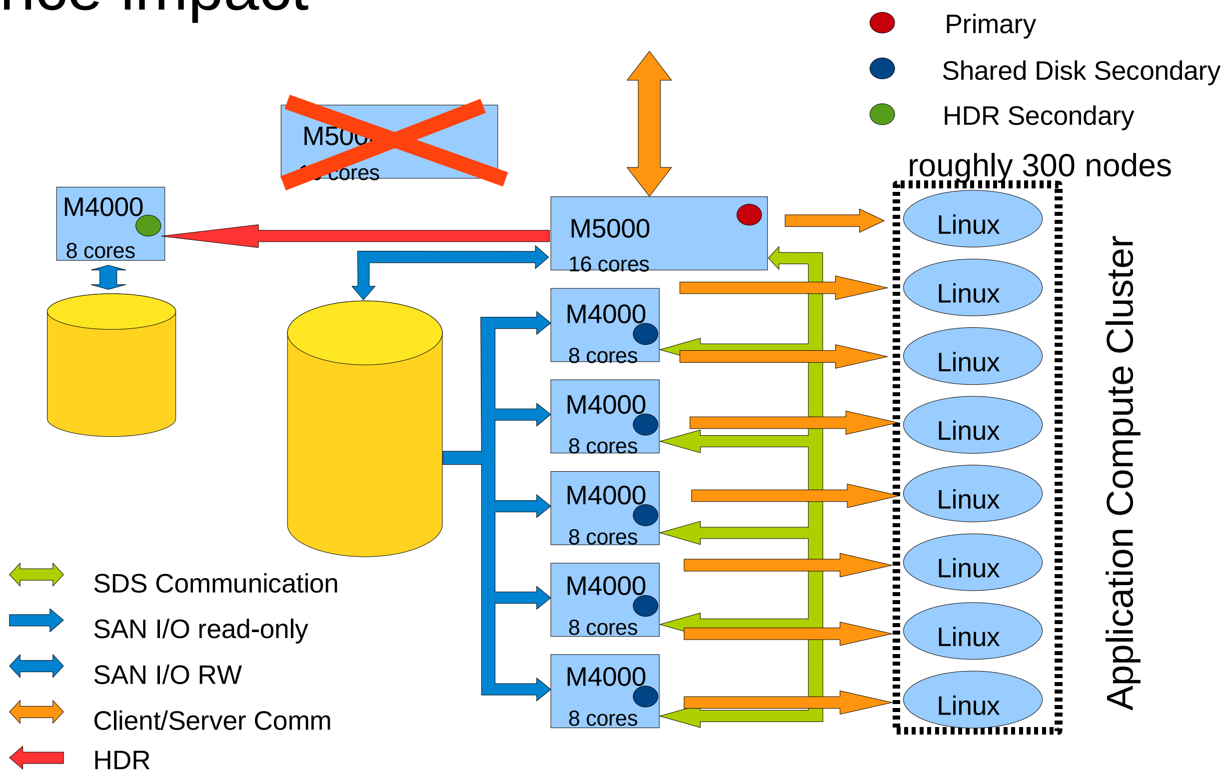
Loss of Primary

- Switch-Over to other M5000 which becomes new primary
- RW I/O access to SAN on new primary
- HDR secondary connects to new primary



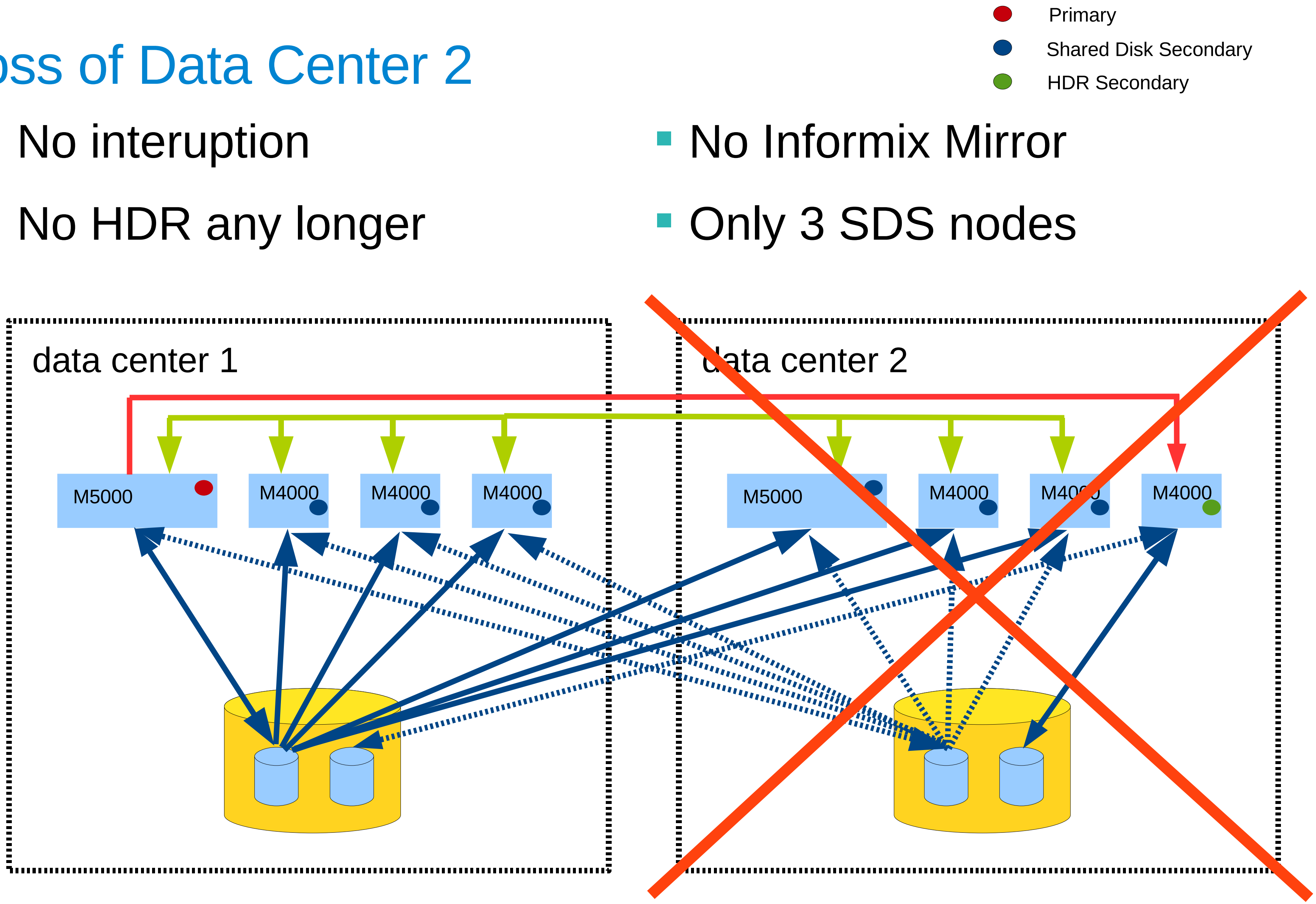
Loss of Primary: After Switch-Over

- RW Clients reconnect to new primary
- Read-Only Clients not affected
- Small performance impact



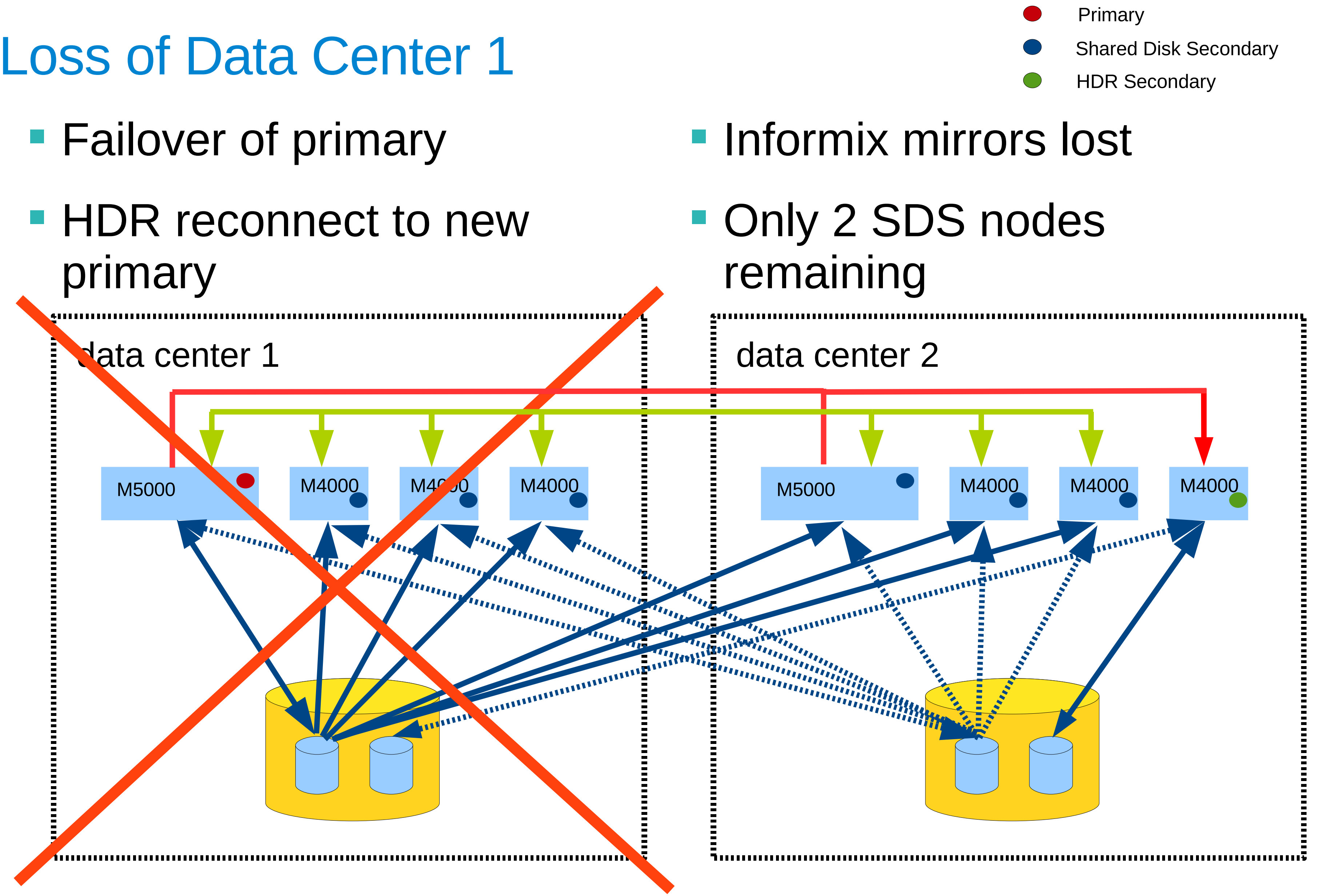
Loss of Data Center 2

- No interuption
- No HDR any longer
- No Informix Mirror
- Only 3 SDS nodes



Loss of Data Center 1

- Failover of primary
- HDR reconnect to new primary
- Informix mirrors lost
- Only 2 SDS nodes remaining



Corruption of Shared Disk

- Assumptions:
  - Problem with dbspace including mirror on SDS Cluster
  - Primary and all shared disk secondaries fail
- Solution:
  - HDR Secondary is only surviving node
  - Becomes Standalone Server
  - Disks still protected by Informix mirroring

Scheduled Maintenance

- Any node may be taken out of the cluster for HW or OS maintenance without interrupting operations
- HDR Secondary:
  - Take out of cluster
  - Maintenance
  - Reconnect and catch up
- Shared Disk Secondary:
  - Take out of cluster
  - Maintenance
  - Reconnect
- Primary:
  - Switch primary to other M5000
  - Maintenance
  - Reconnect as Shared Disk Secondary
  - Optional: Switch primaries again

Migration

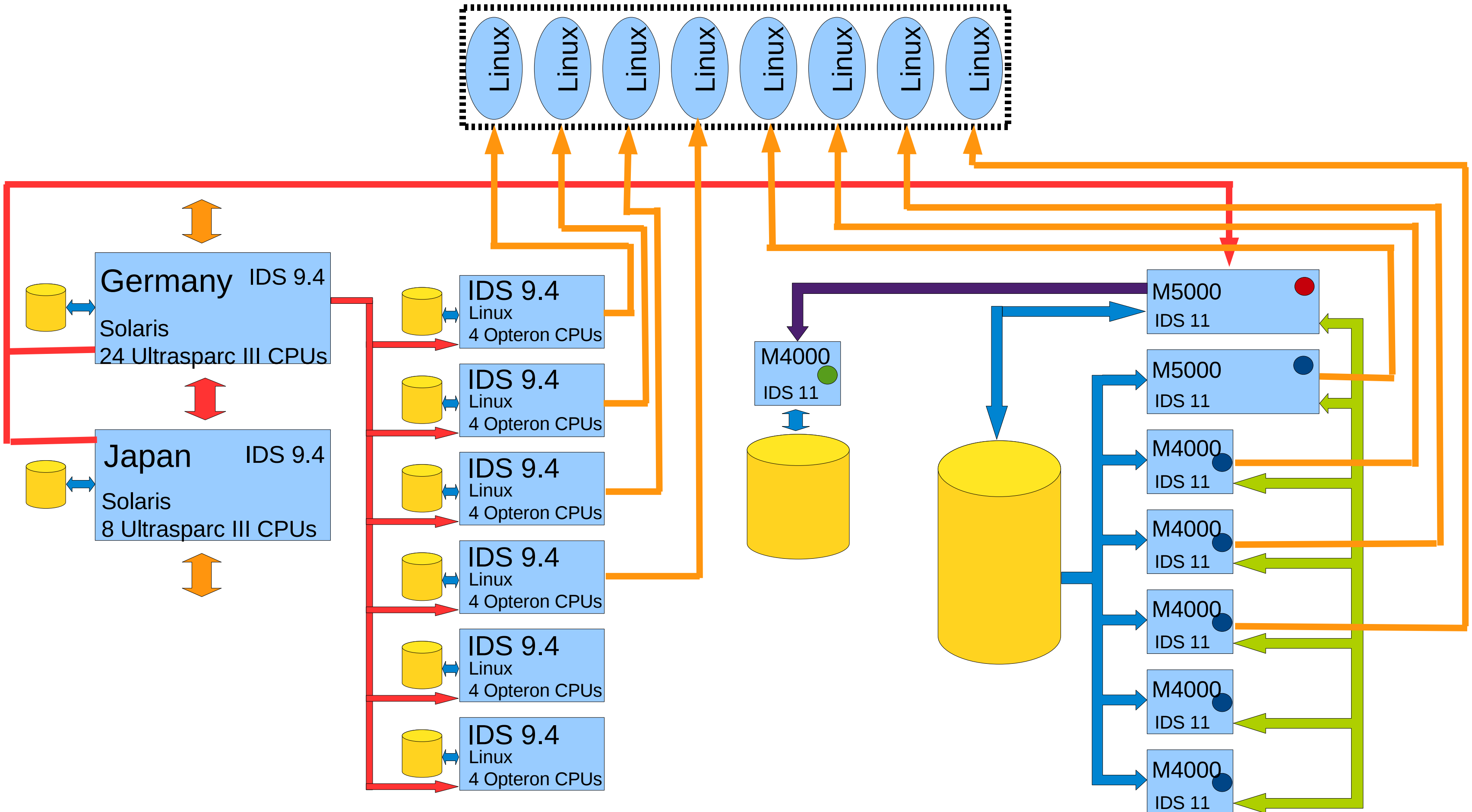
Requirements for Migration to New System

- “Smooth“ migration:
  - No system outage
  - No performance degradation during migration
  - Fast switch to old system in case of problems
  - Parallel operating of old and new system
- Just 3 months from planning to going live
- No or only minimal application changes
- Project started before general availability of IDS 11.50

Implementation and Test

- Single node tests of IDS 11 on M4000:
  - Functionality
  - Performance
- Shared Disk Test: Primary with one Shared Disk Secondary
  - Performance
  - Availability
  - Flow control
- Combined Shared Disk and HDR Test
- Integration Test of all nodes

Parallel Operation of Old and New System



Disk I/O

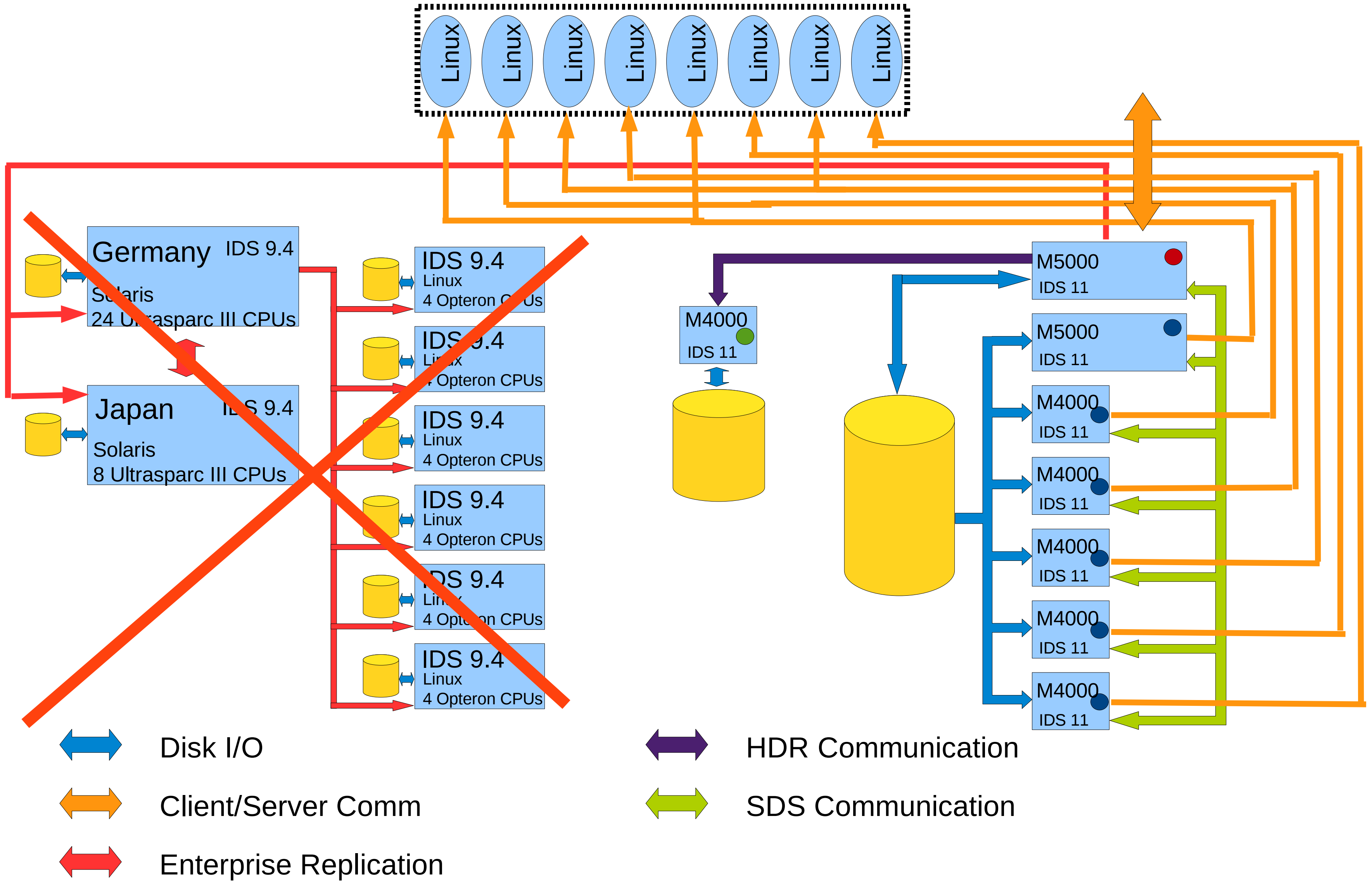
Client/Server Comm

Enterprise Replication

HDR Communication

SDS Communication

Migration Final Step



Lessons Learned

Experiences

- Significant cost reduction
- Mach 11 cluster without any 3rd party software possible
- Less than 3 month from 1st planning steps to start of production
- No major problems uncovered during intensive testing
- Easily extensible with additional SDS nodes (scalability)
- High degree of availability
- Fast disaster recovery

Thanks!