



A Database Architecture for Scalability and High Availability

Andreas Weinger, IBM

Andreas.Weinger@de.ibm.com





Agenda

- Trends in DBMS Requirements
- MACH 11 Technology
 - Shared Disk Secondaries (SDS)
 - HDR/RSS
 - Enterprise Replication
- Example
 - Problem
 - Design of the MACH 11 Cluster
 - Performance
 - High Availability
- Conclusions



Trends in DBMS Requirements





Trends in DBMS Requirements: Availability

- Many applications may be accessed from anywhere in the world at any time
 - Application/Database must be available 24/7
 - No unplanned downtime
 - No planned downtime
 - No Maintenance Windows any longer available
 - Upgrade of HW, OS, DBMS during regular operation



Trends in DBMS Requirements: Scalability

- Flexible reaction to workload changes
- Growing by addition of small, cheap computers
- Scale-up and scale-out



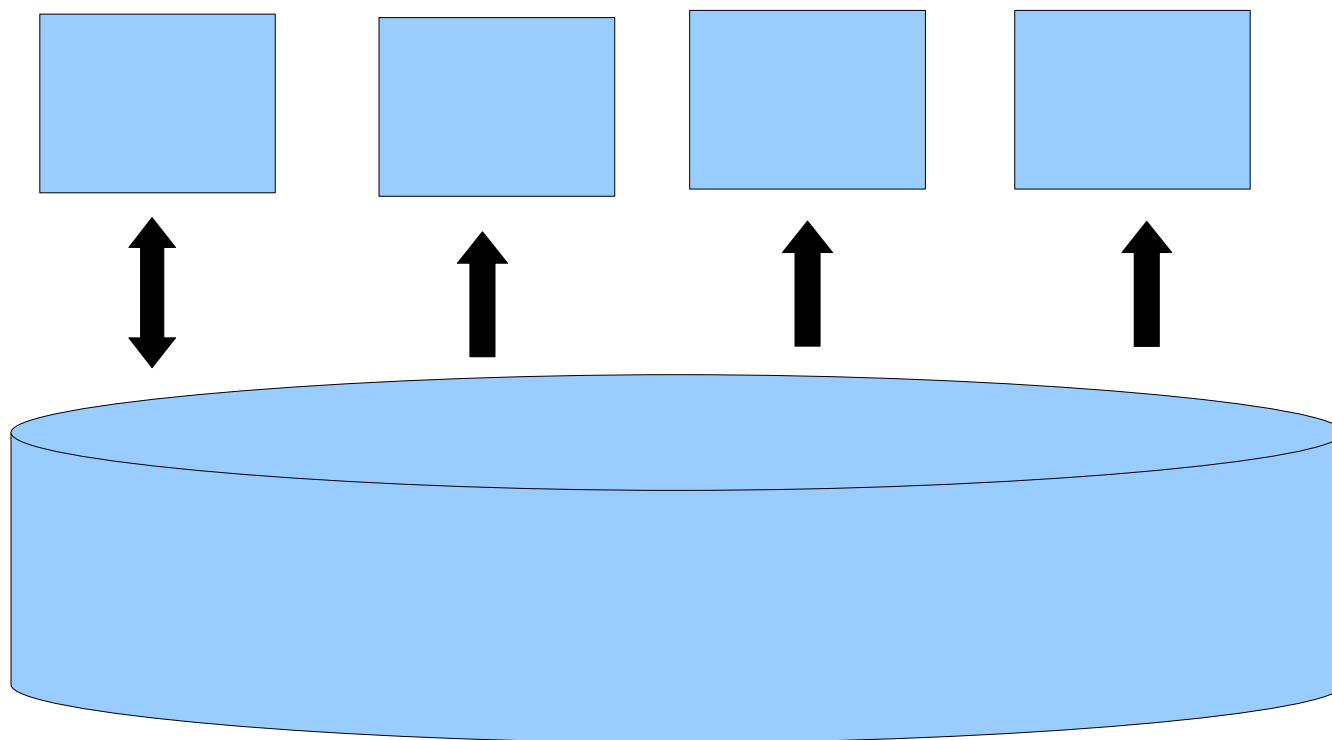
MACH 11 Technology

Multi-node Active Cluster for High Availability



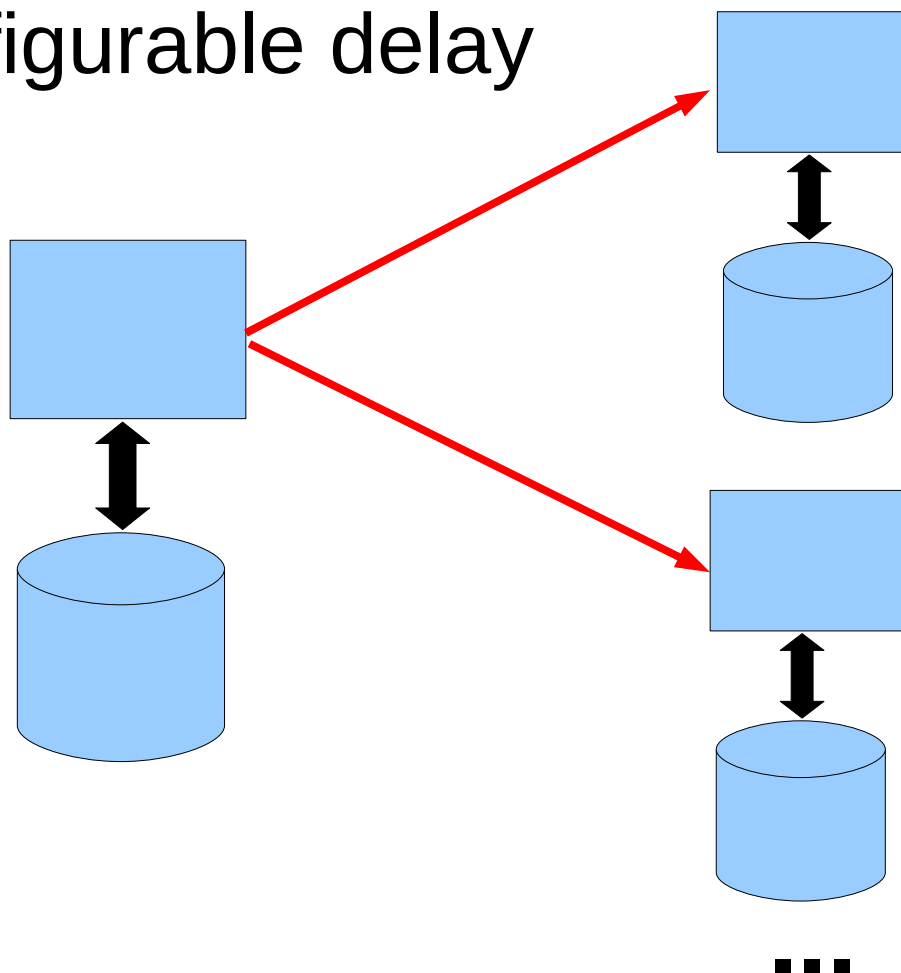
Shared Disk Secondaries (SDS)

- Shared Disk Technology
- Primary synchronizes disk access



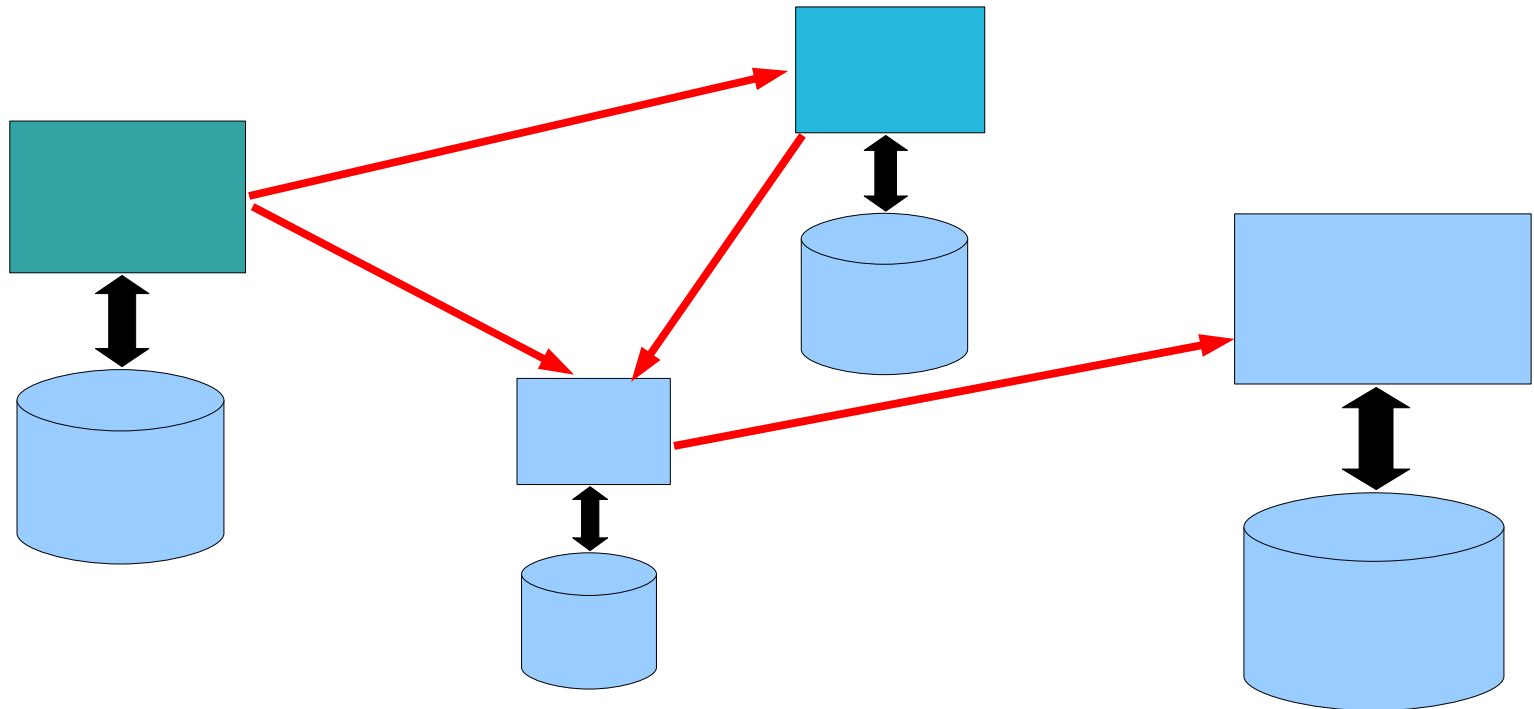
HDR/RSS

- Log Record Replication (sync / async)
- configurable delay



Enterprise Replication

- Heterogeneous Replication
 - HW, OS, IDS version
- Rolling Upgrade of all components possible





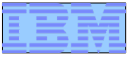
Example





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 the securities (access is read-only)
- New information about securities permanently inserted
- Application is very business critical; therefore high availability requirements



Requirements for Database Architecture

- high availability
- fast disaster recovery
- good scalability
- performance
- Smooth transition from old system
- low cost

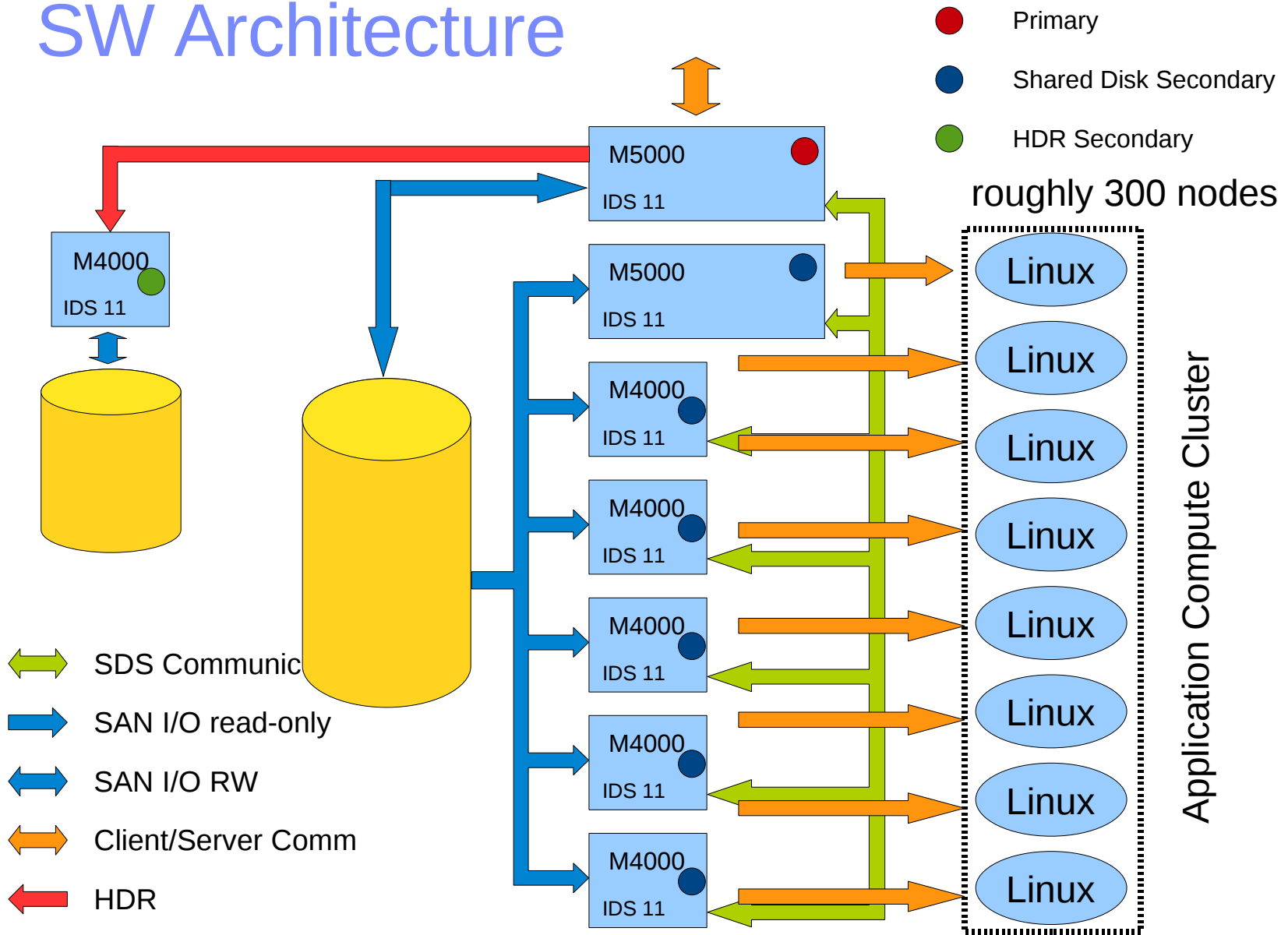


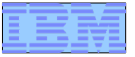
Design Mach 11 Cluster



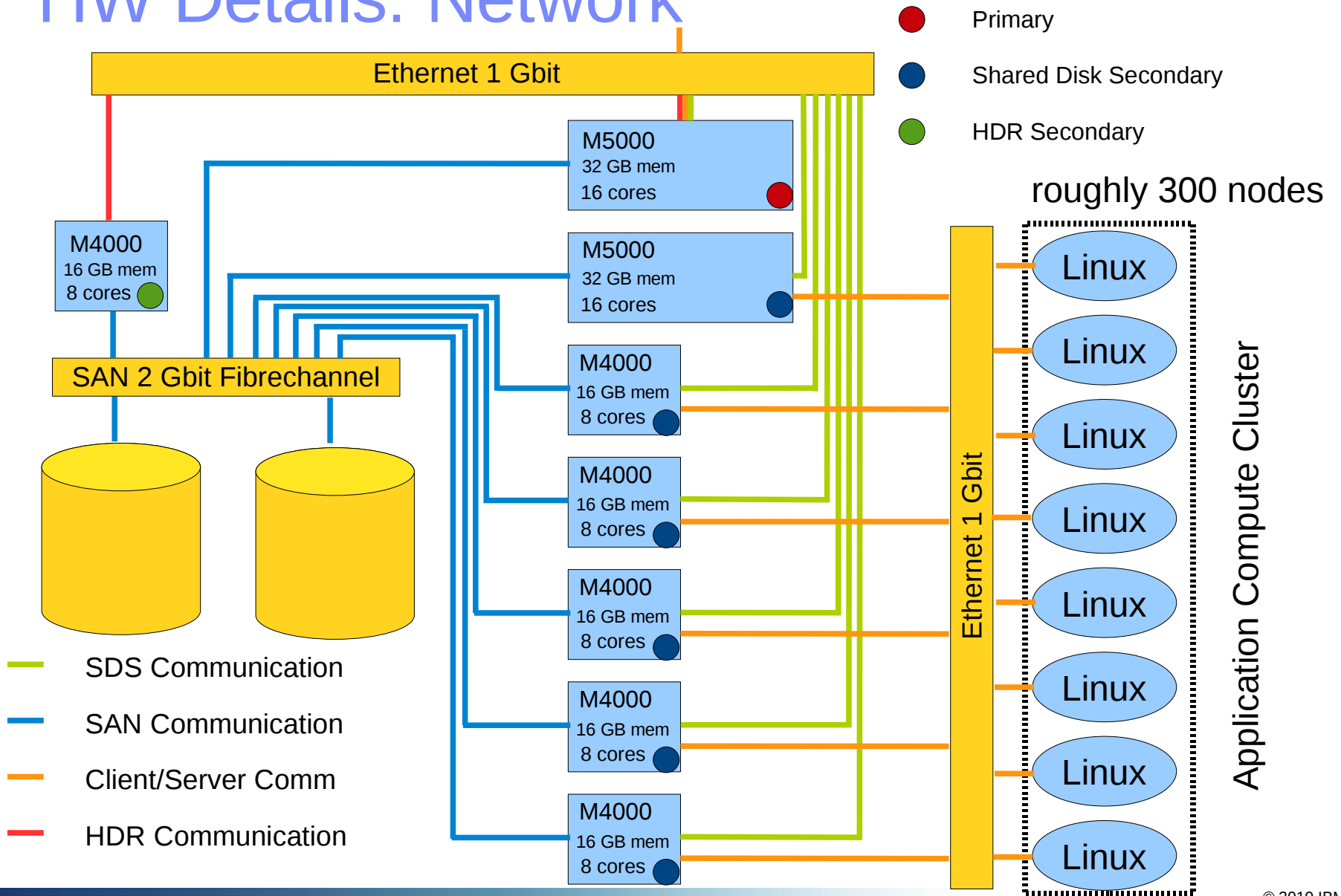


SW Architecture





HW Details: Network



Application Compute Cluster



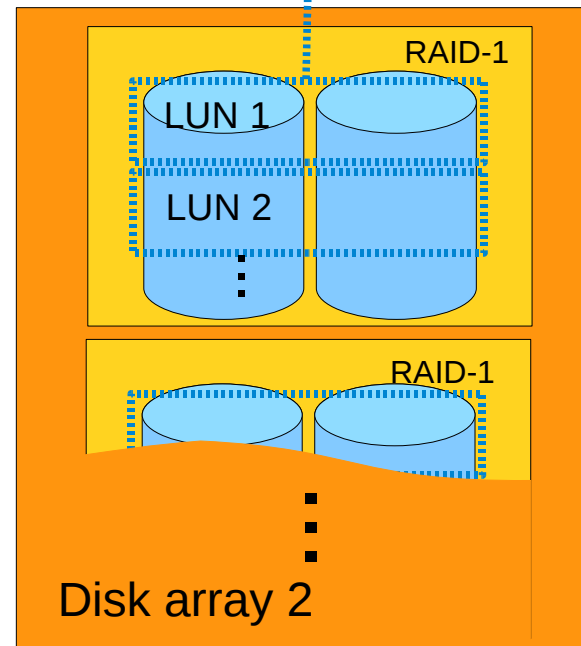
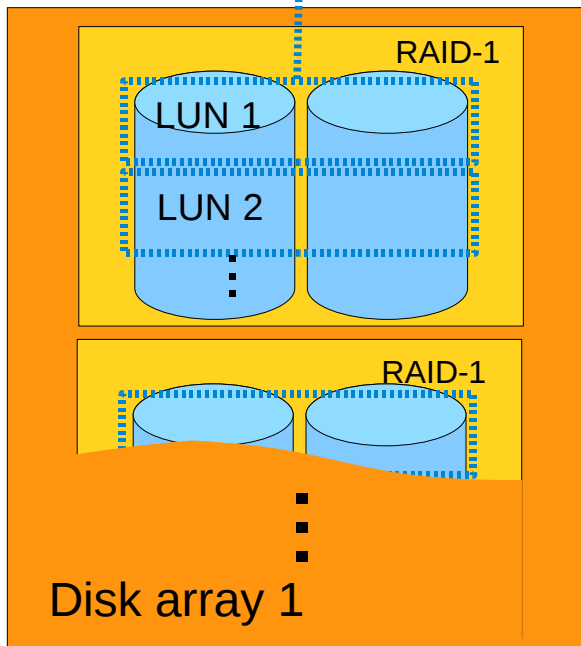
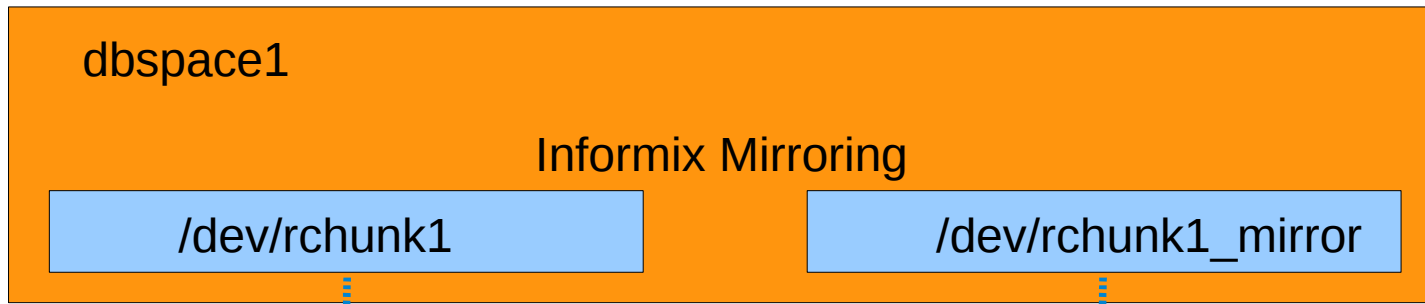
Redundancy at the Disk Level

- 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





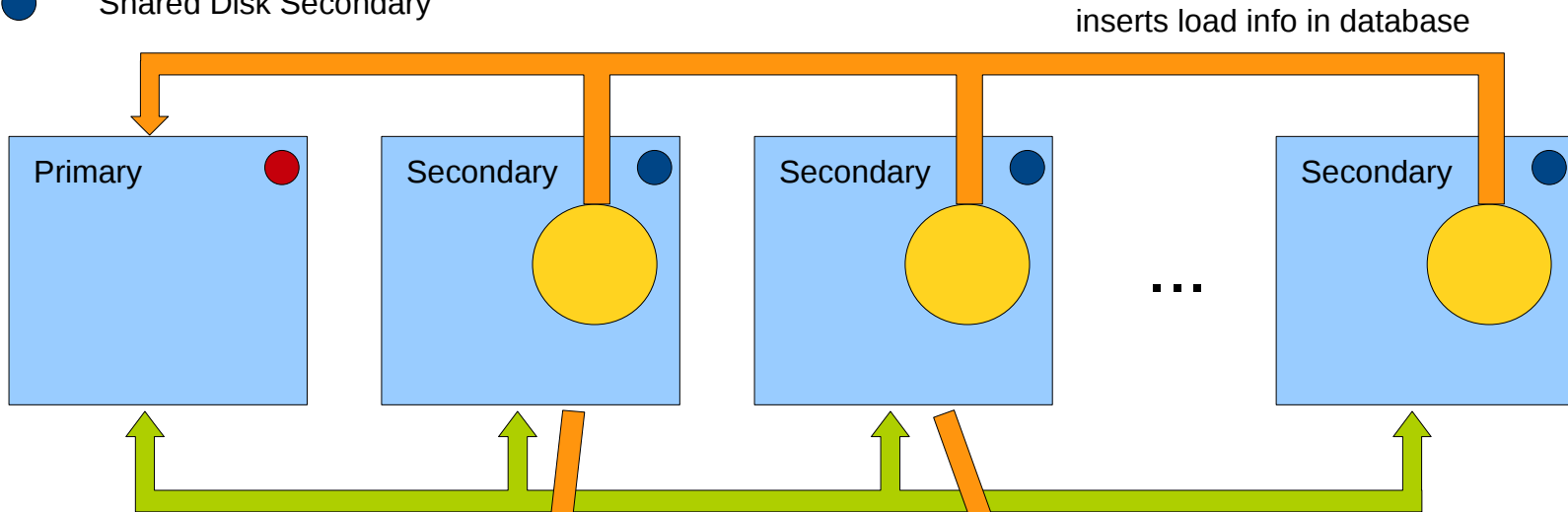
Why was the mapping of the dbspaces done this way?

- 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

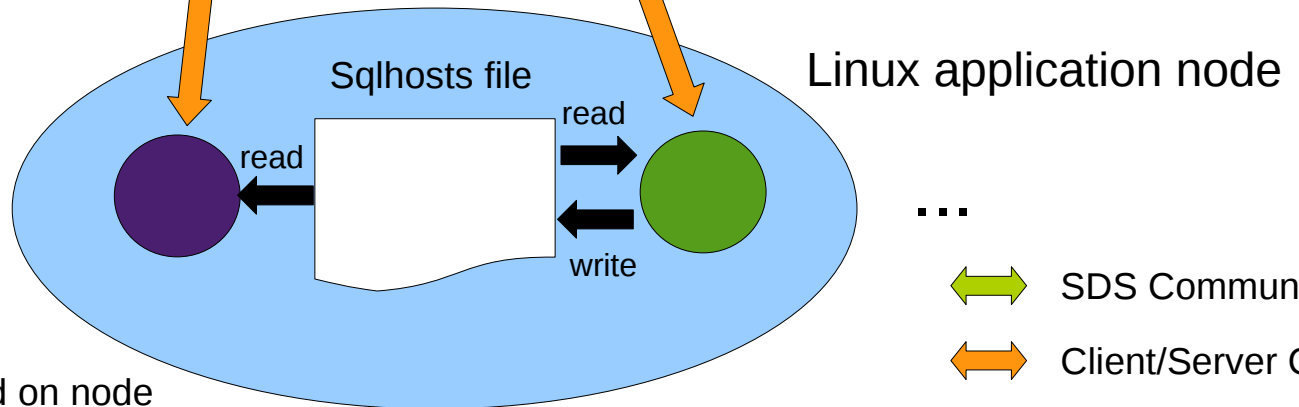


Communication with the Compute Cluster

- Primary
- Shared Disk Secondary



- application client
- "load balancer"
- Monitor: measures load on node

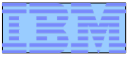


- ↔ SDS Communication
- ↔ Client/Server Comm



Performance





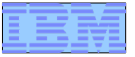
Avoiding Bottlenecks

- 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



Performance Measurements

- Log throughput without secondary for stress test:
 - 1 GB / 25 sec
- Log throughput with HDR secondary for stress test:
 - 1 GB / 2 min
- Log throughput with all SDS nodes stay the same since SDS much faster than HDR
- HDR used to prevent diverging of shared disk primary and secondaries



High Availability



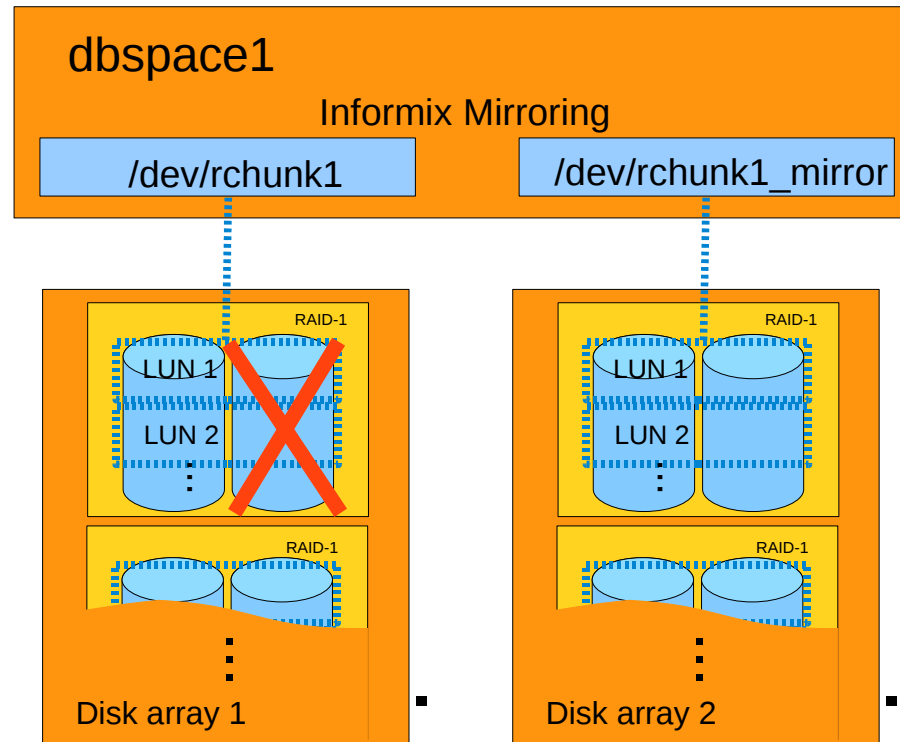


Different Availability Scenarios

- 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

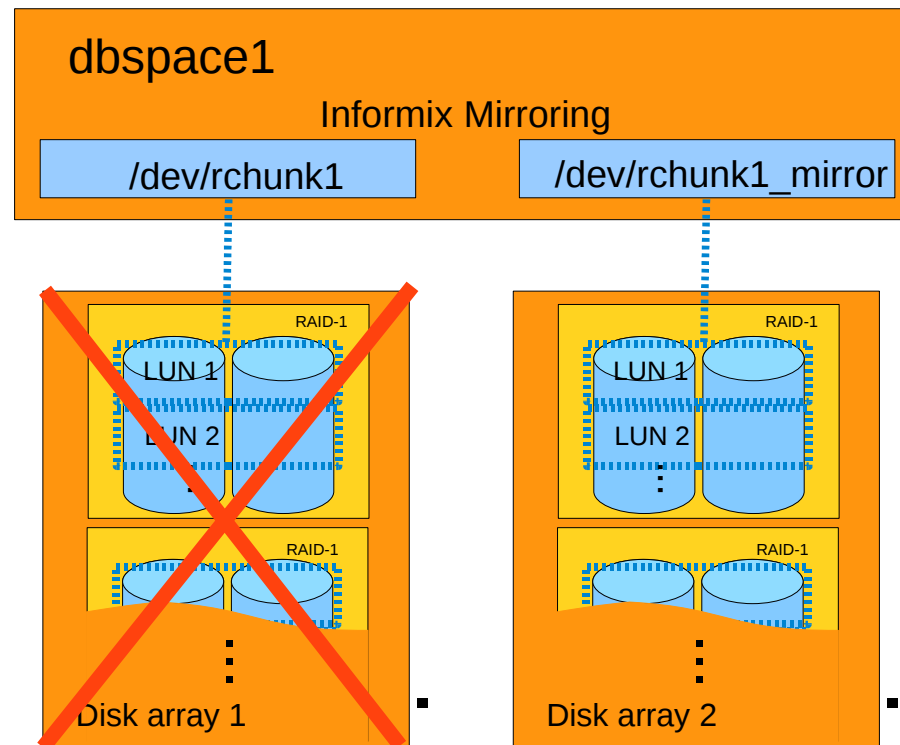
Loss of a physical disk

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



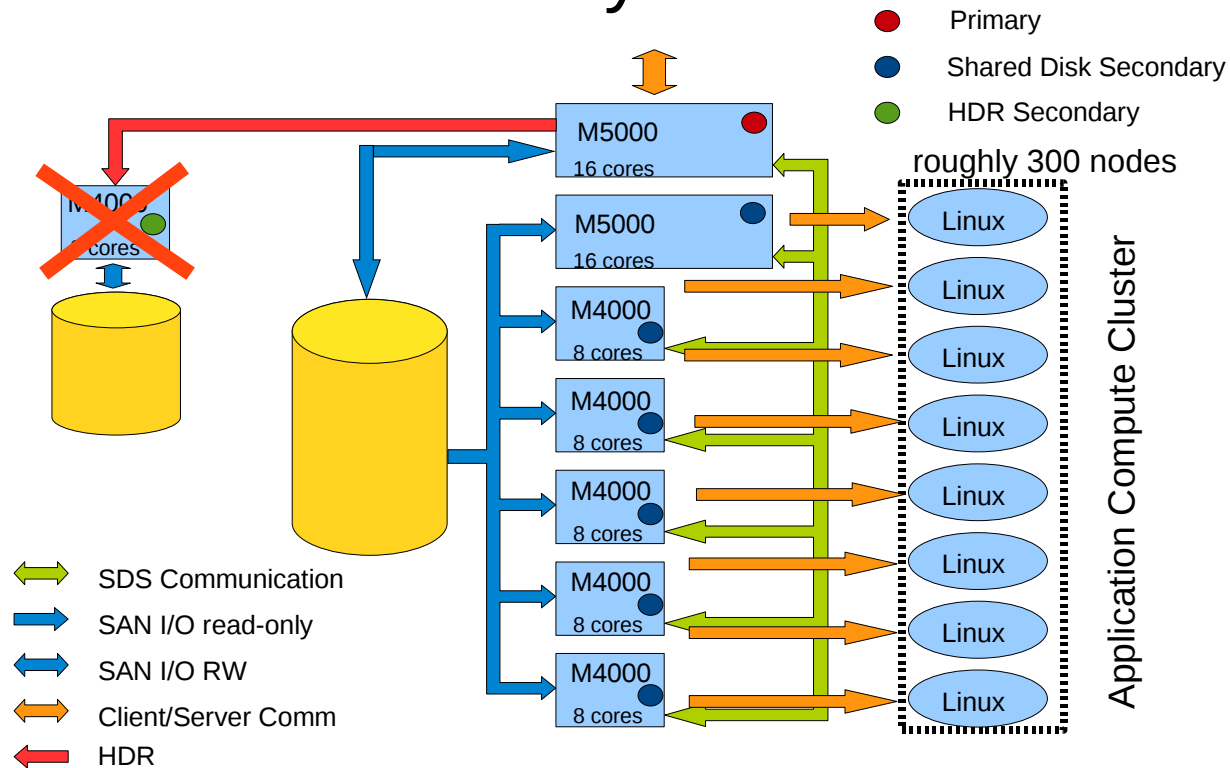
Loss of a disk array

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



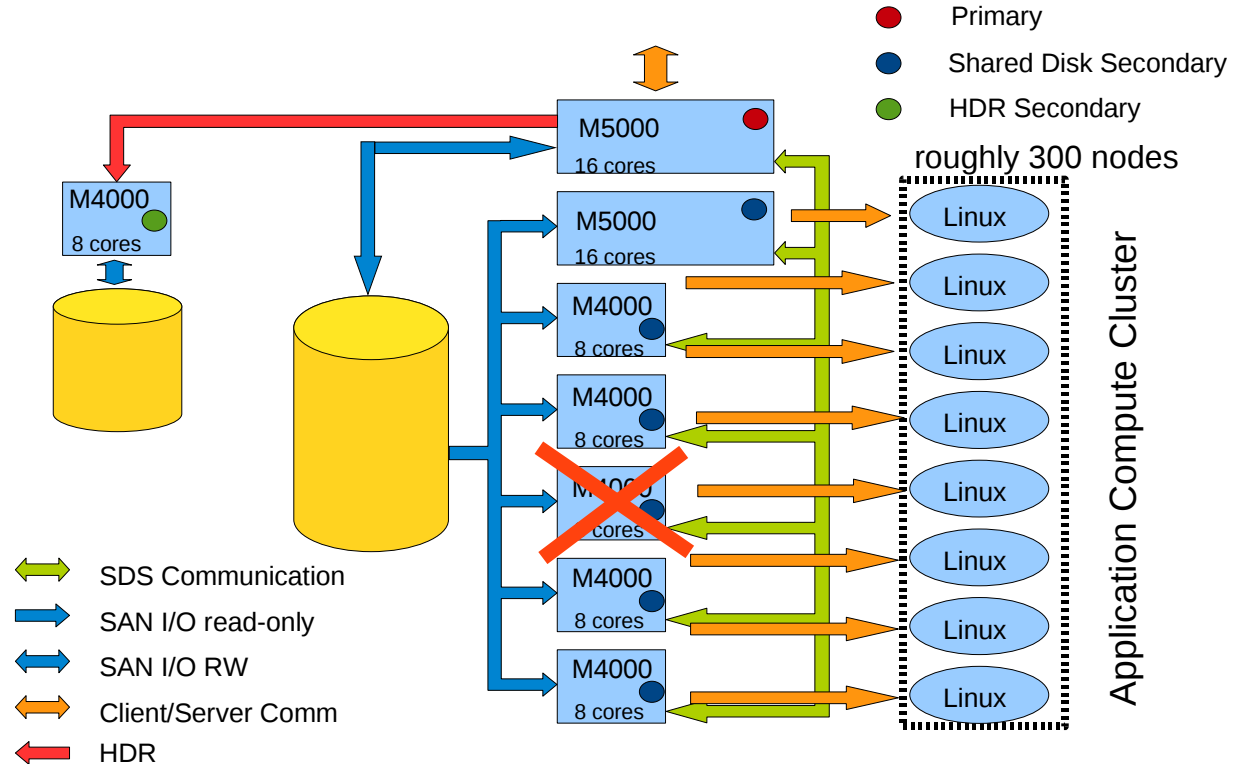
Loss of HDR Secondary

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



Loss of Shared Disk Secondary

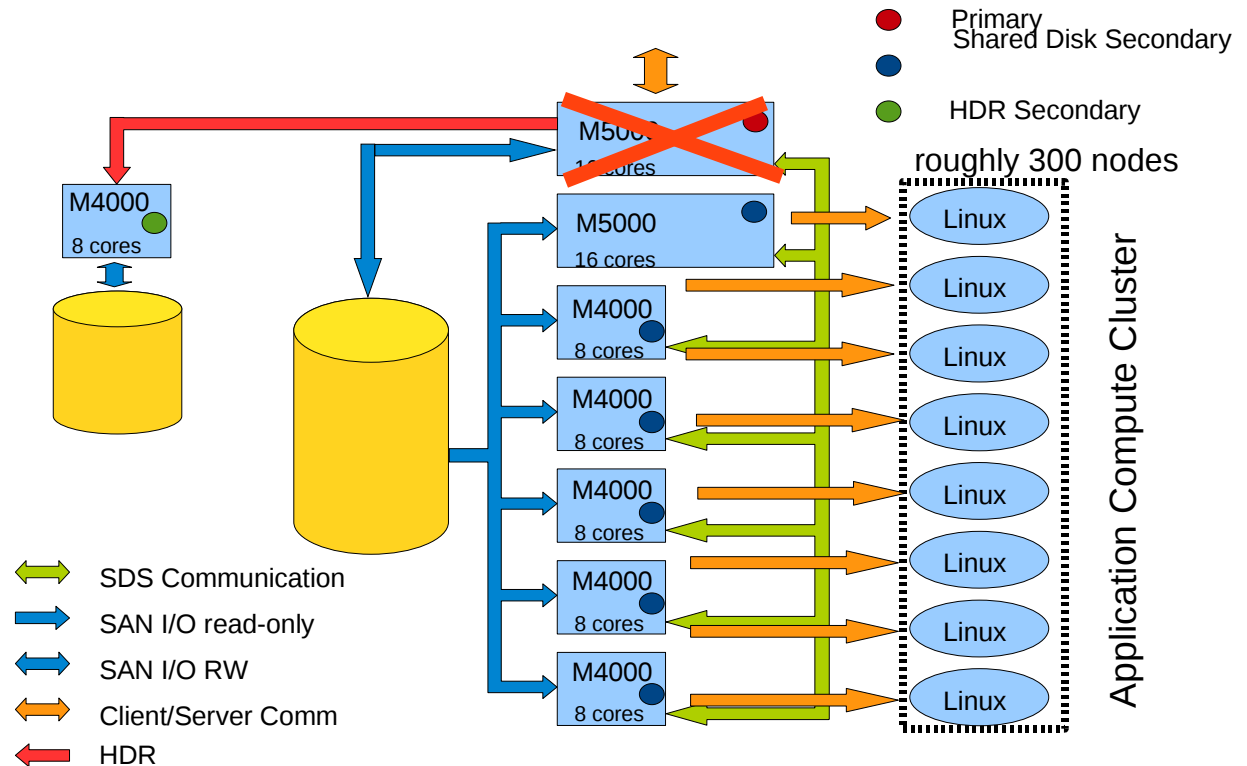
- No immediate impact on remaining servers
- Clients of failed node have to reconnect to other SDS node
- Node no longer used by load balance
- Performance impact





Loss of Primary

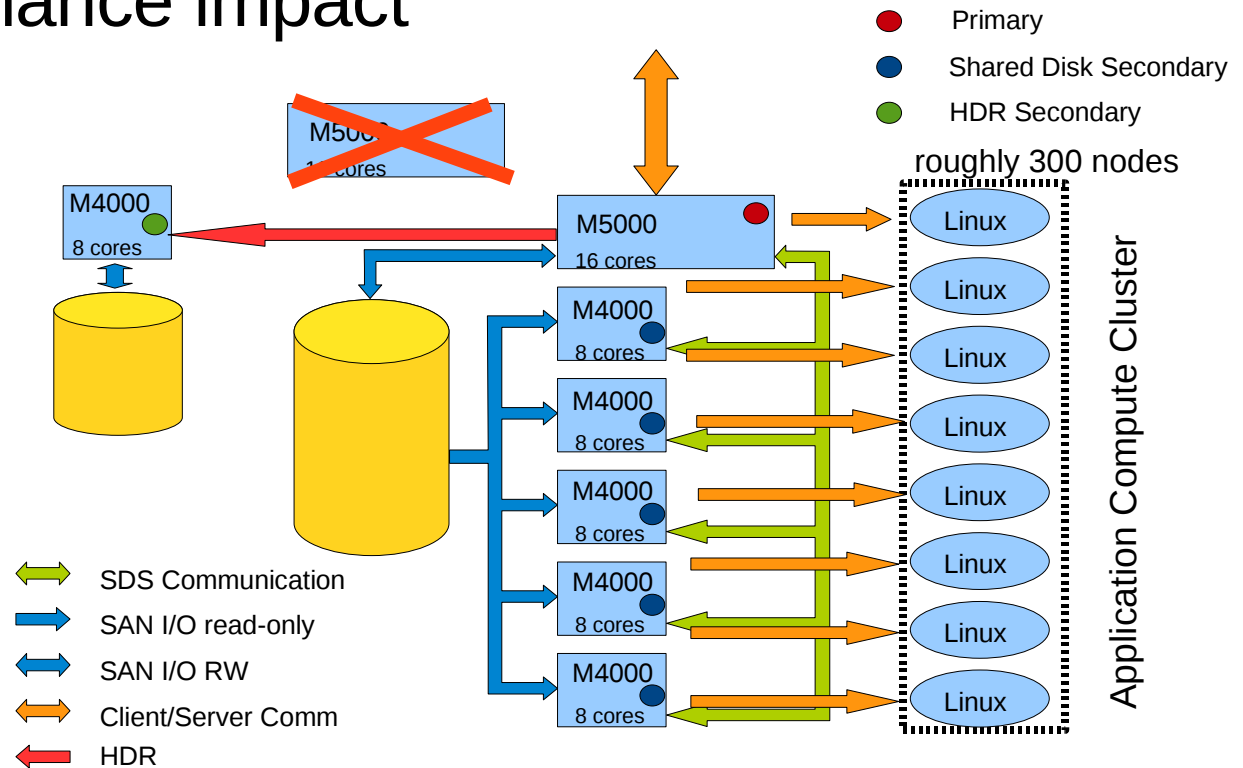
- Switch-Over to other M5000 which becomes new primary
- RW I/O access to SAN on new primary
- HDR secondary connects to new primary





Loss of Primary: After Switch-Over

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

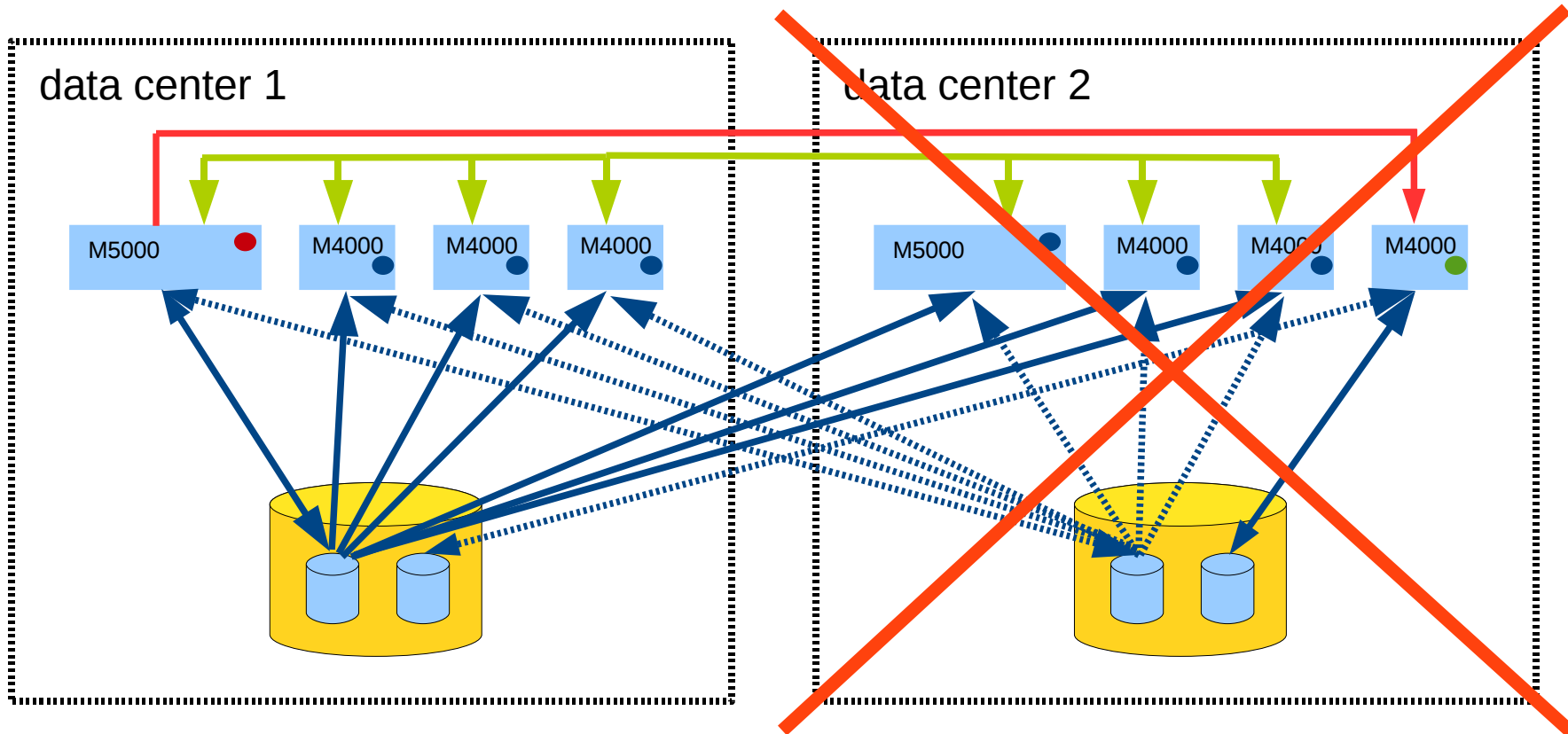




- Primary
- Shared Disk Secondary
- HDR Secondary

Loss of Data Center 2

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



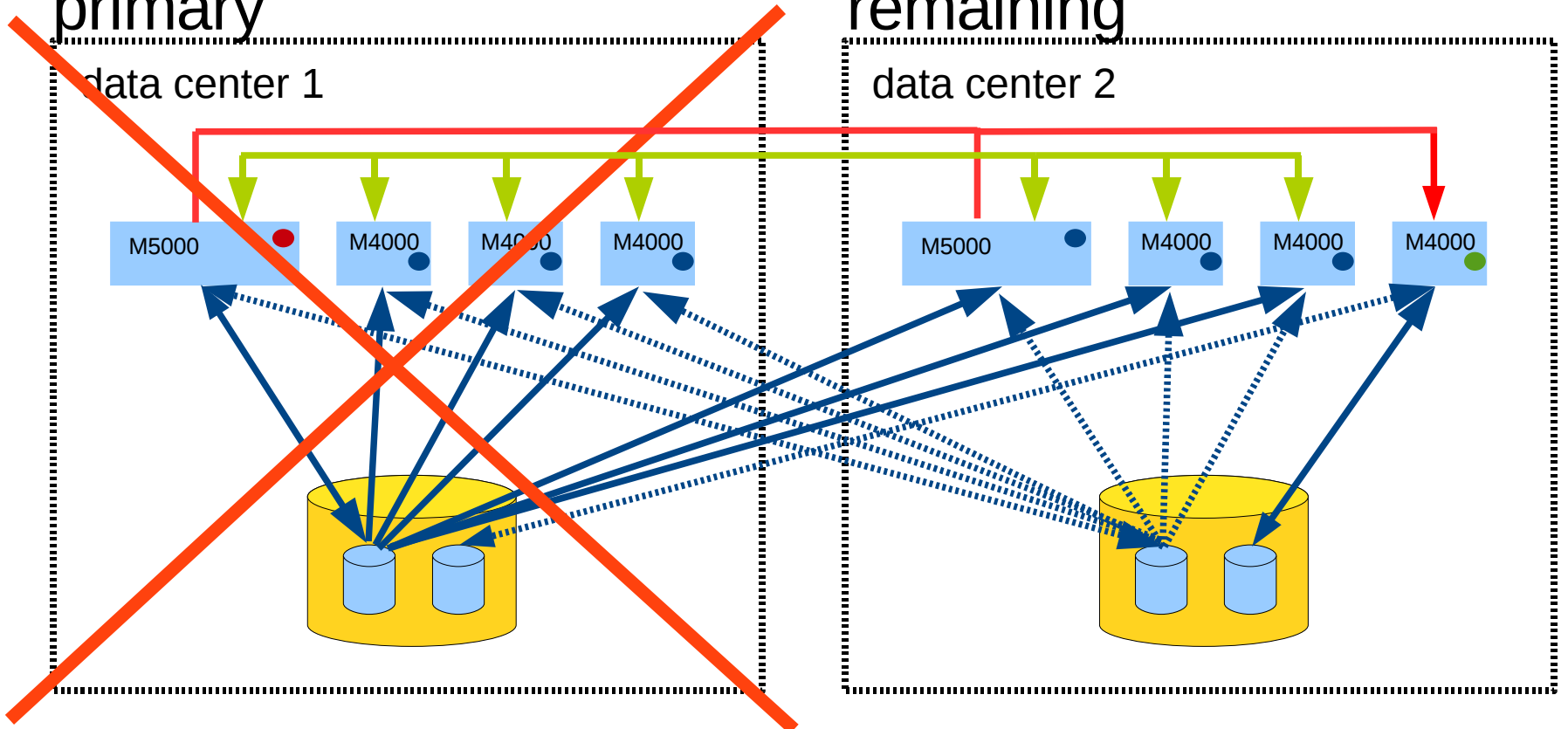


- Primary
- Shared Disk Secondary
- HDR Secondary

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

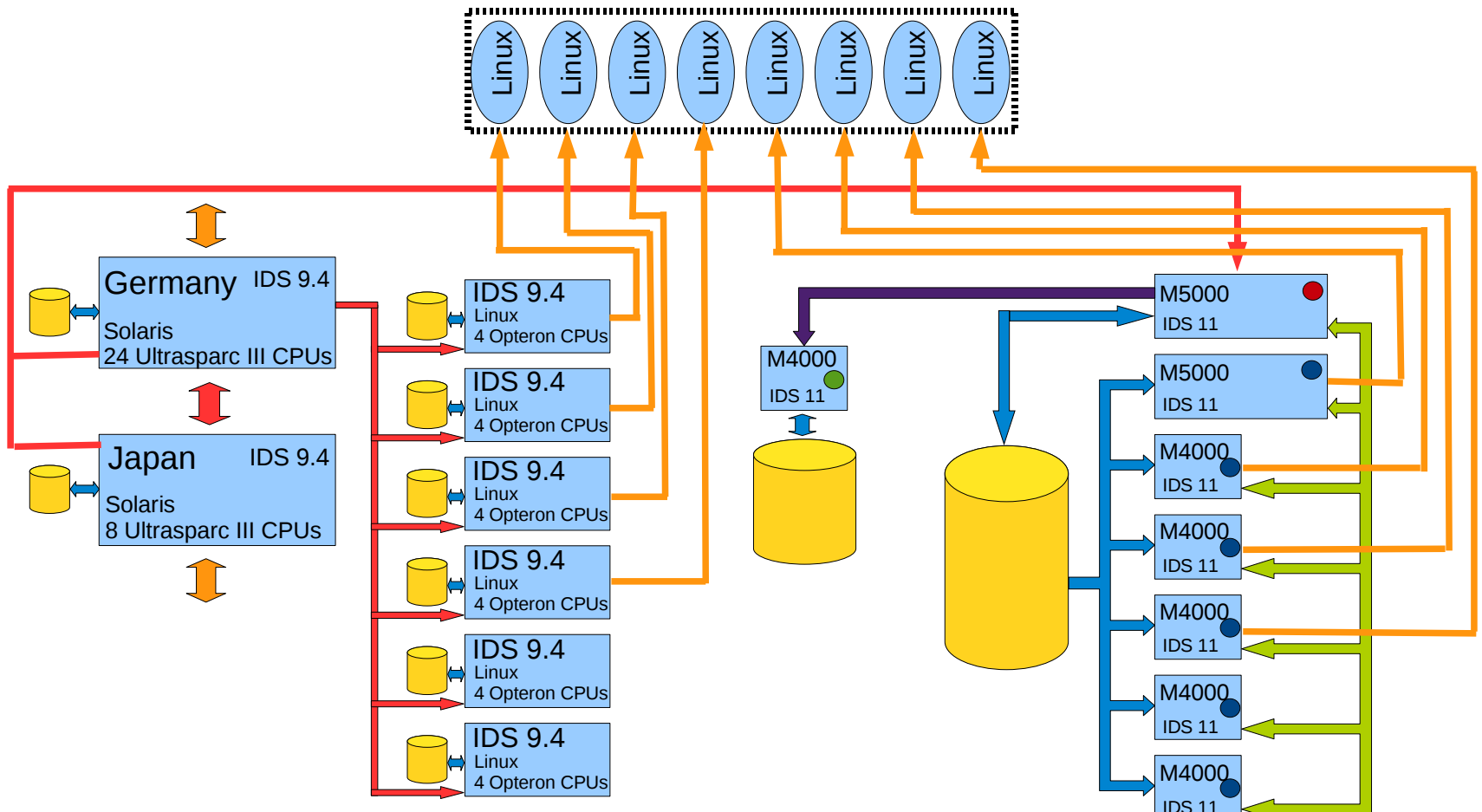


Scheduled Maintenance

- 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



Parallel Operation of Old and New System

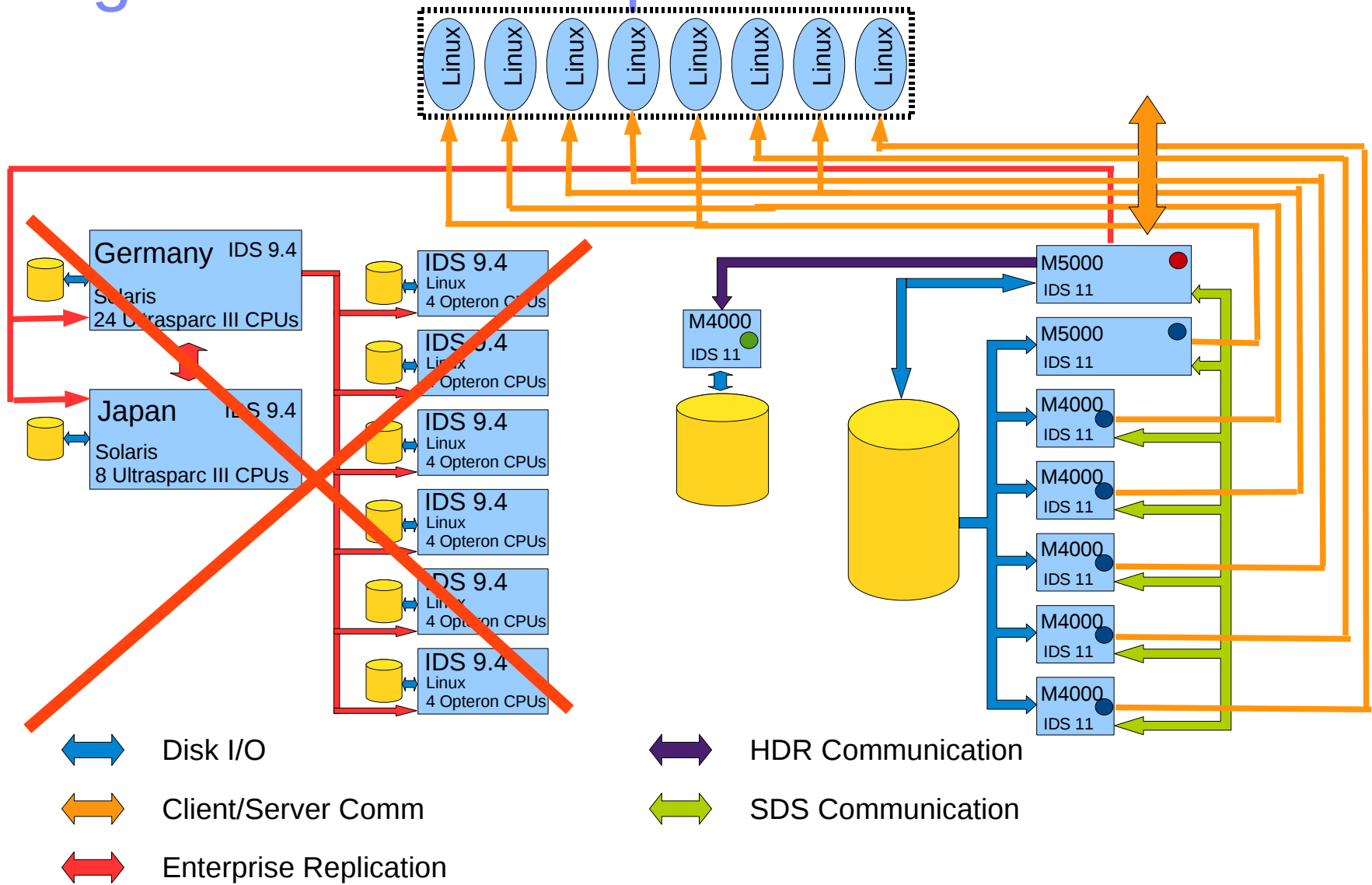


- Disk I/O
- Client/Server Comm
- Enterprise Replication

- HDR Communication
- SDS Communication



Migration Final Step





Conclusions





Conclusions

- Database architecture should be designed for individual requirements of each project
- MACH 11 provides functionality to implement diverse availability and scalability requirements
- Standard HW components sufficient for providing extreme high degree of availability and very good scalability
- SDS nodes provide mechanism to adapt easily and fast to changing workload requirements



Thank you!

