Early Experiences with CXFS and Storage Area Networks

John Lynch
(303)581-4451
John.k.lynch@lmco.com
25 May, 2001

Agenda

AEROJET

A GenCorp Company

- Introduction
- SAN Design Requirements
- Final SAN Architecture
- SAN Performance
  - Processor Utilization
  - Metadata Network
- Trials and Tribulations
  - Network
  - Two Node Cluster
  - Filesystem Layout
- Installed System Photo Layout
- Future Work
- Summary
- Questions

Introduction

AEROJET

A GenCorp Company

- Aerojet is a government contractor
- Industry leader in providing
  - Missile warning data processing ground stations
  - Space-based Infared optical sensors
  - Propulsion Systems
  - Smart Munitions

SAN Design Requirements

AEROJET

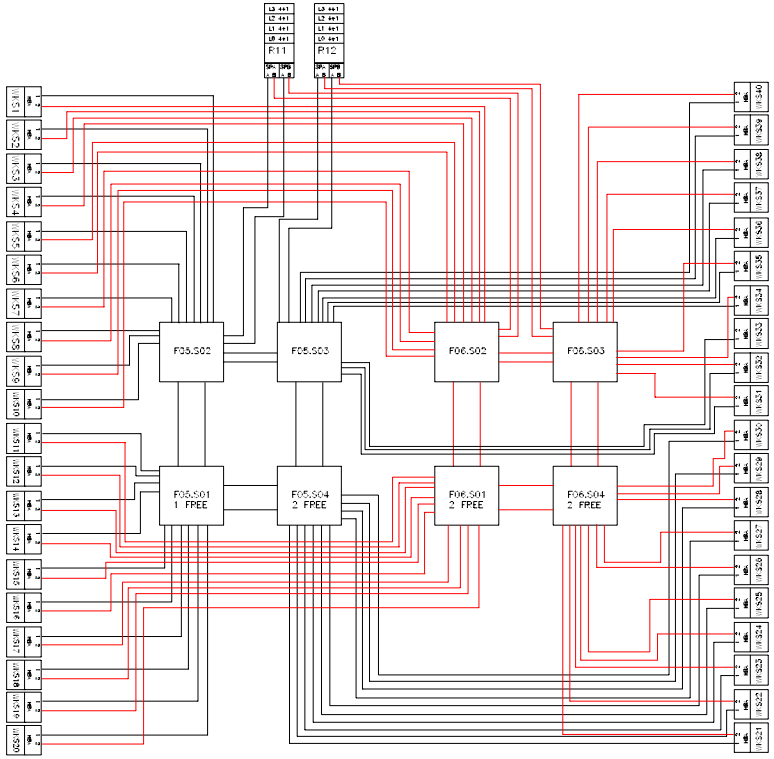
A GenCorp Company

- SAN access speed from server to disk comparable with direct attach Ultra SCSI
- No single point of failure
- Single failure in SAN fabric does not create system failure
- Metadata network compatible with existing ethernet infrastructure
- Economically justifiable
- Compatible with existing tape storage devices
- 100 percent CPU and 150 percent network margins maintained

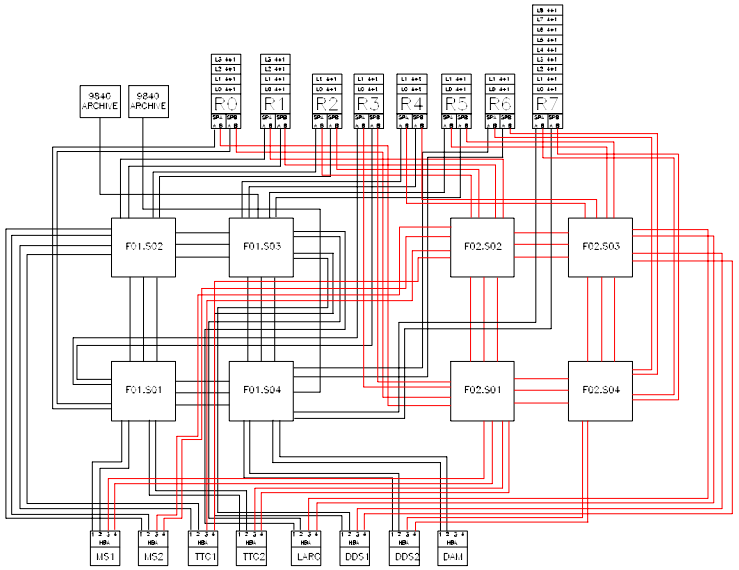
Final SAN Architecture

AEROJET

A GenCorp Company

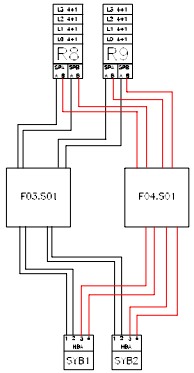


Fabric 3



Fabric 1

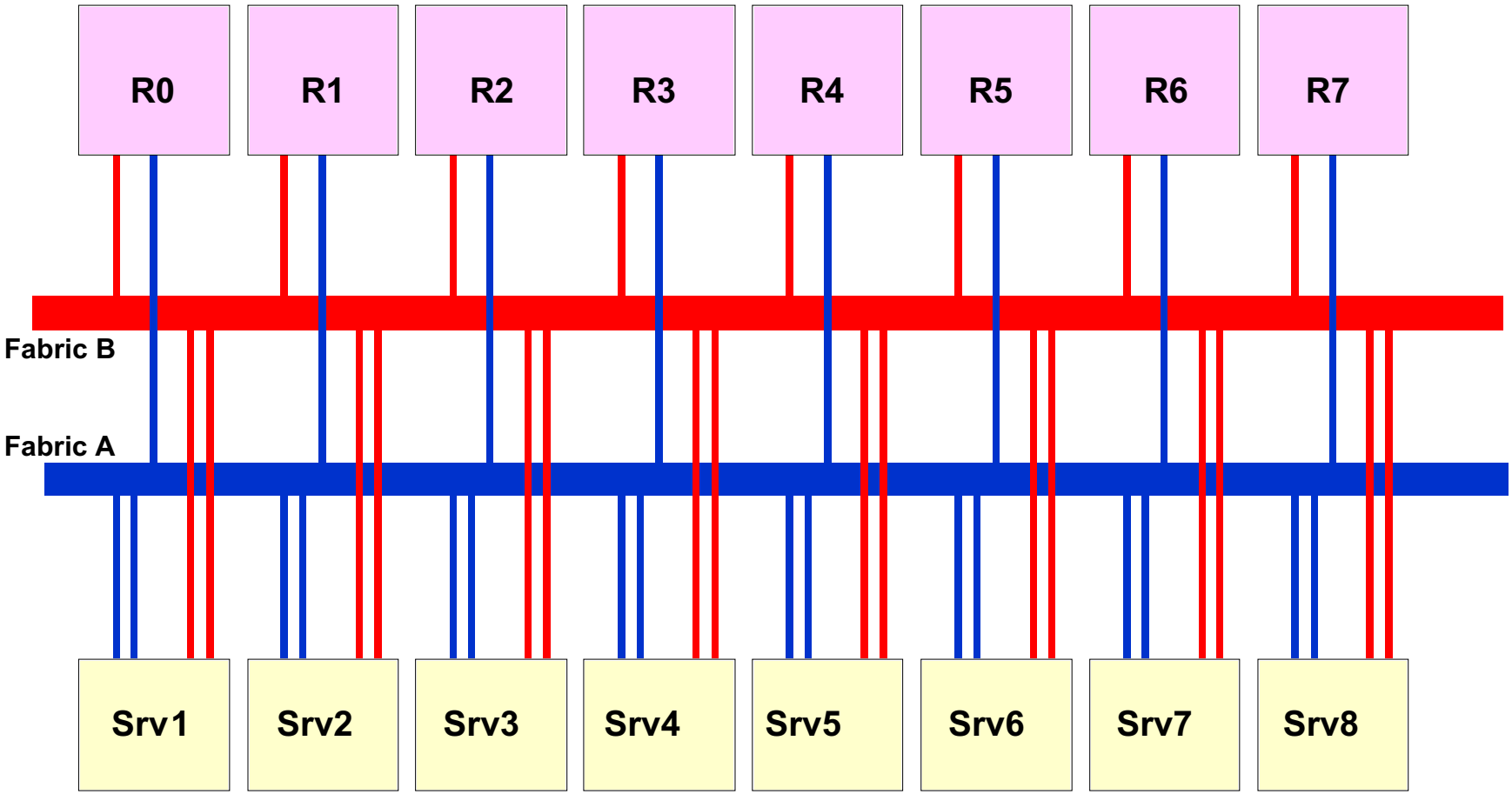
Fabrics 2&4



SAN Fabric 1

AEROJET

A GenCorp Company

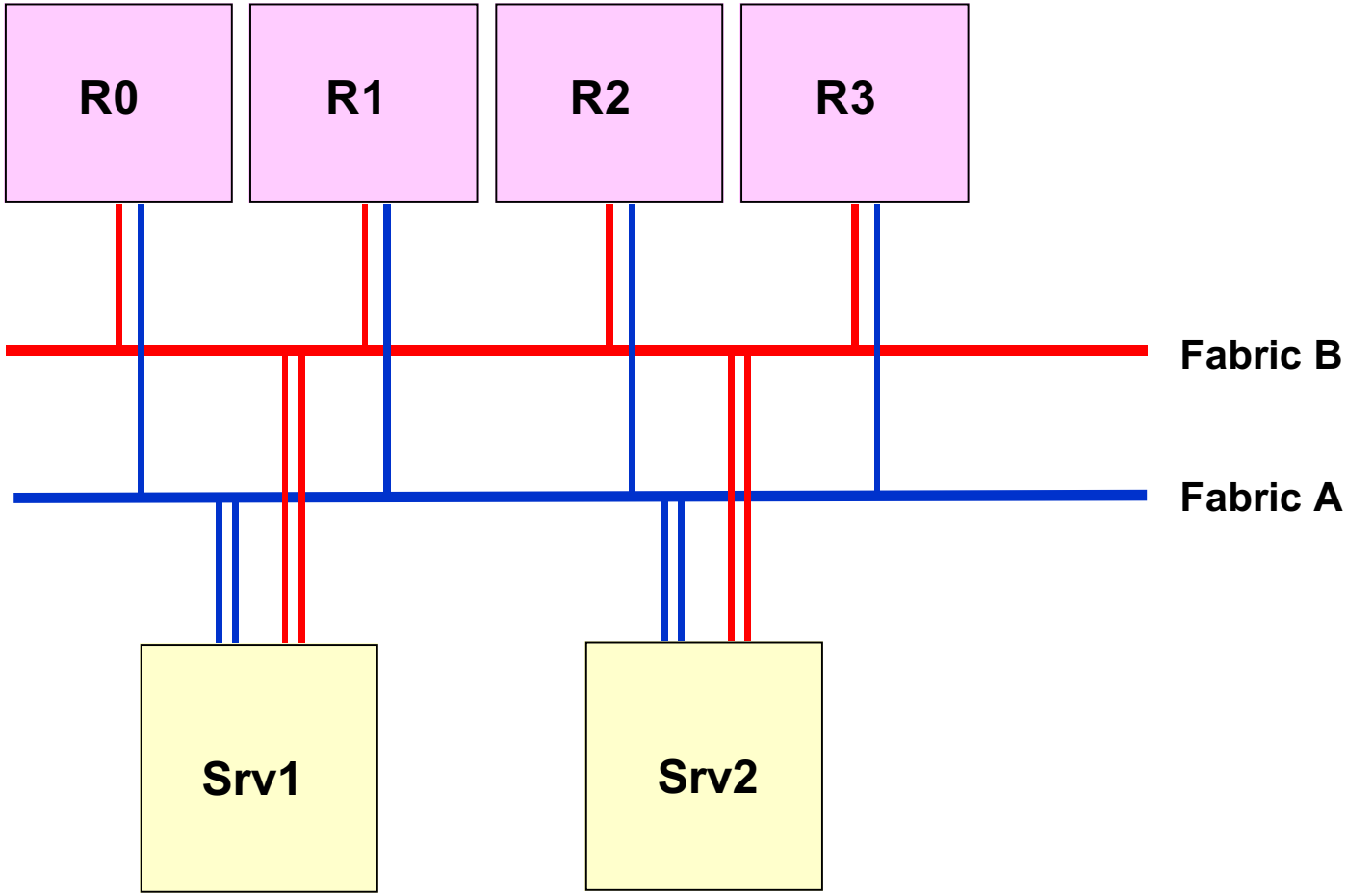


Total Storage Capacity 2.3TB

SAN Fabric 2

AEROJET

A GenCorp Company

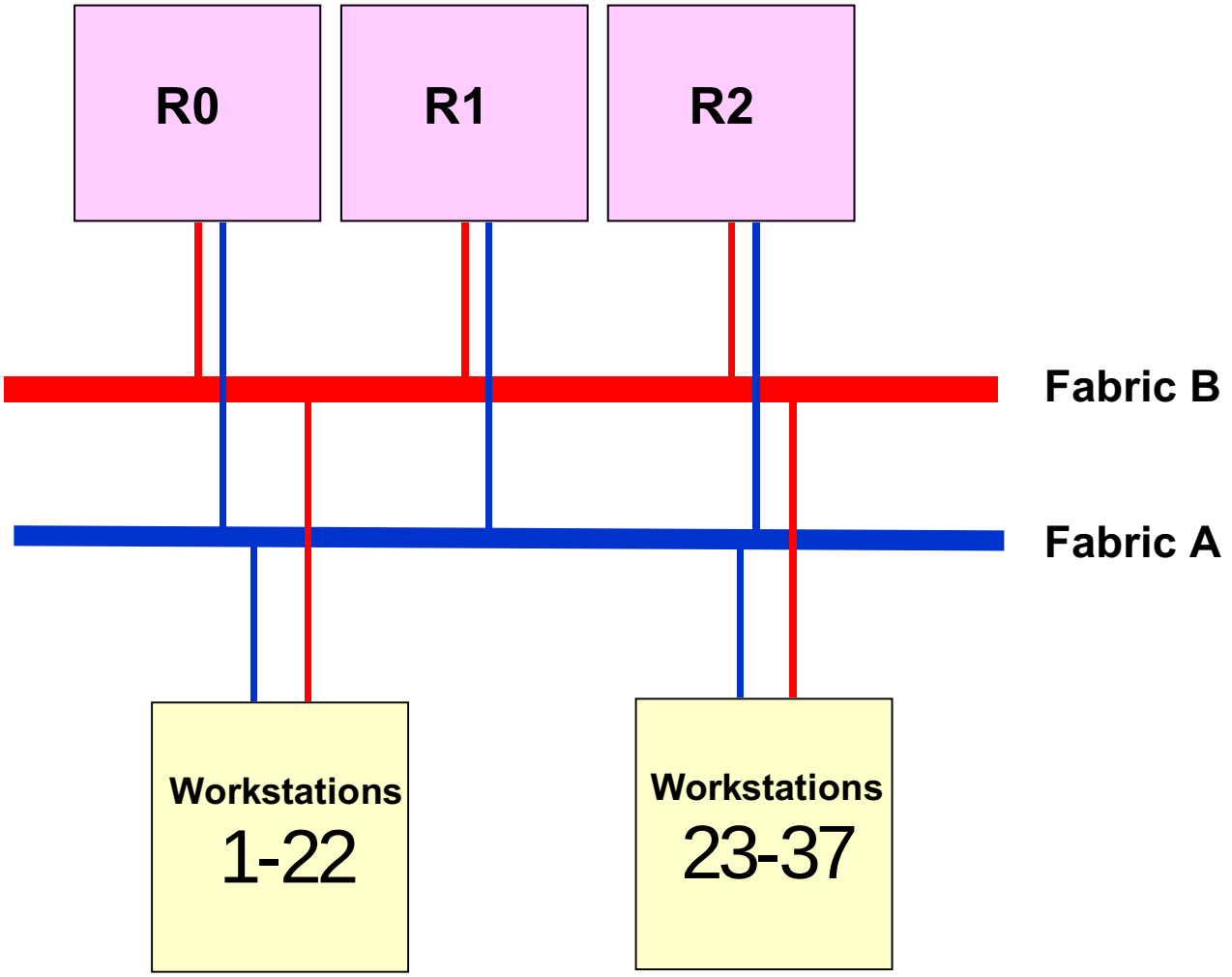


Total Storage Capacity 512MB

SAN Fabric 3

AEROJET

A GenCorp Company

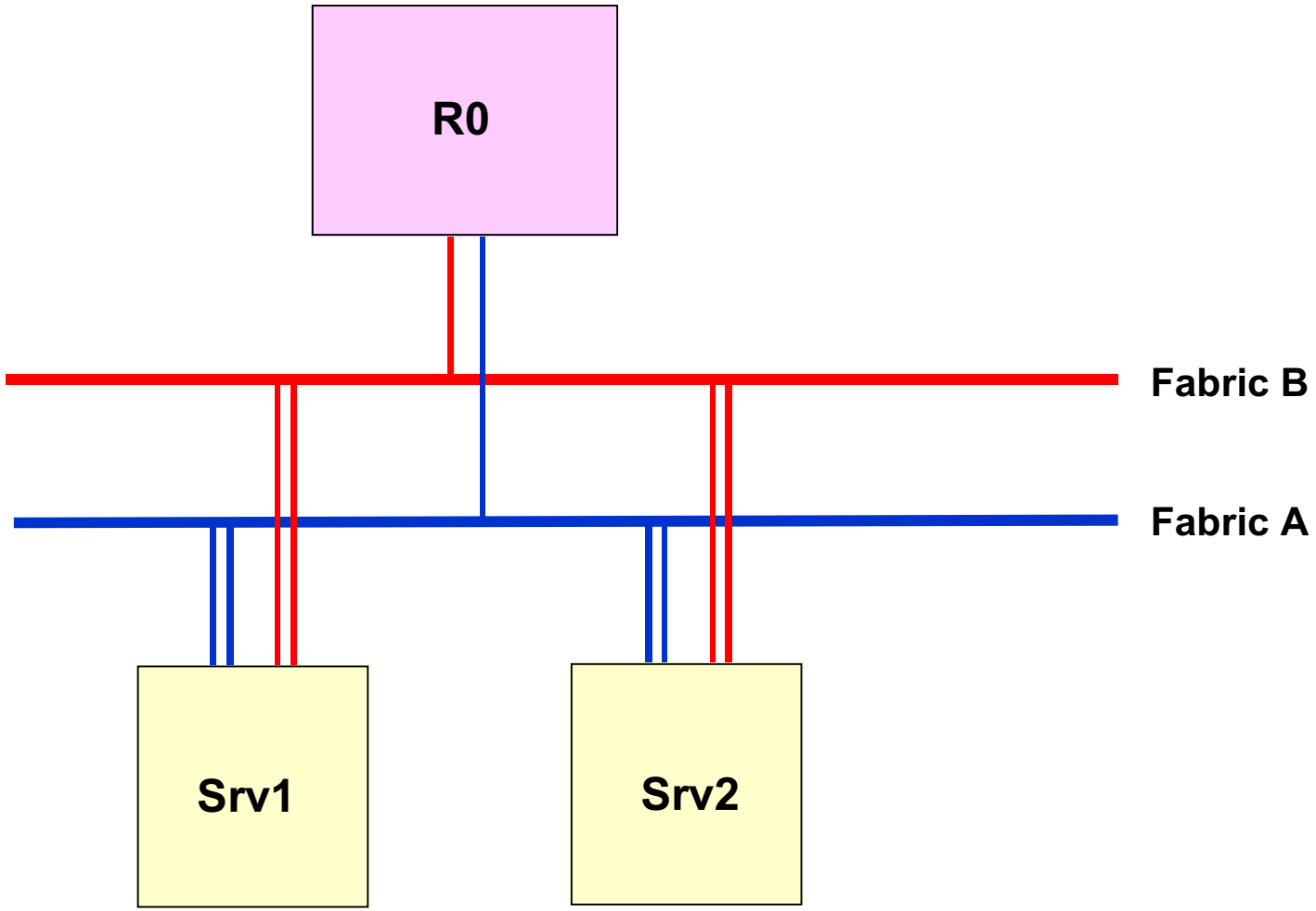


Total Storage Capacity 384MB

SAN Fabric 4

AEROJET

A GenCorp Company

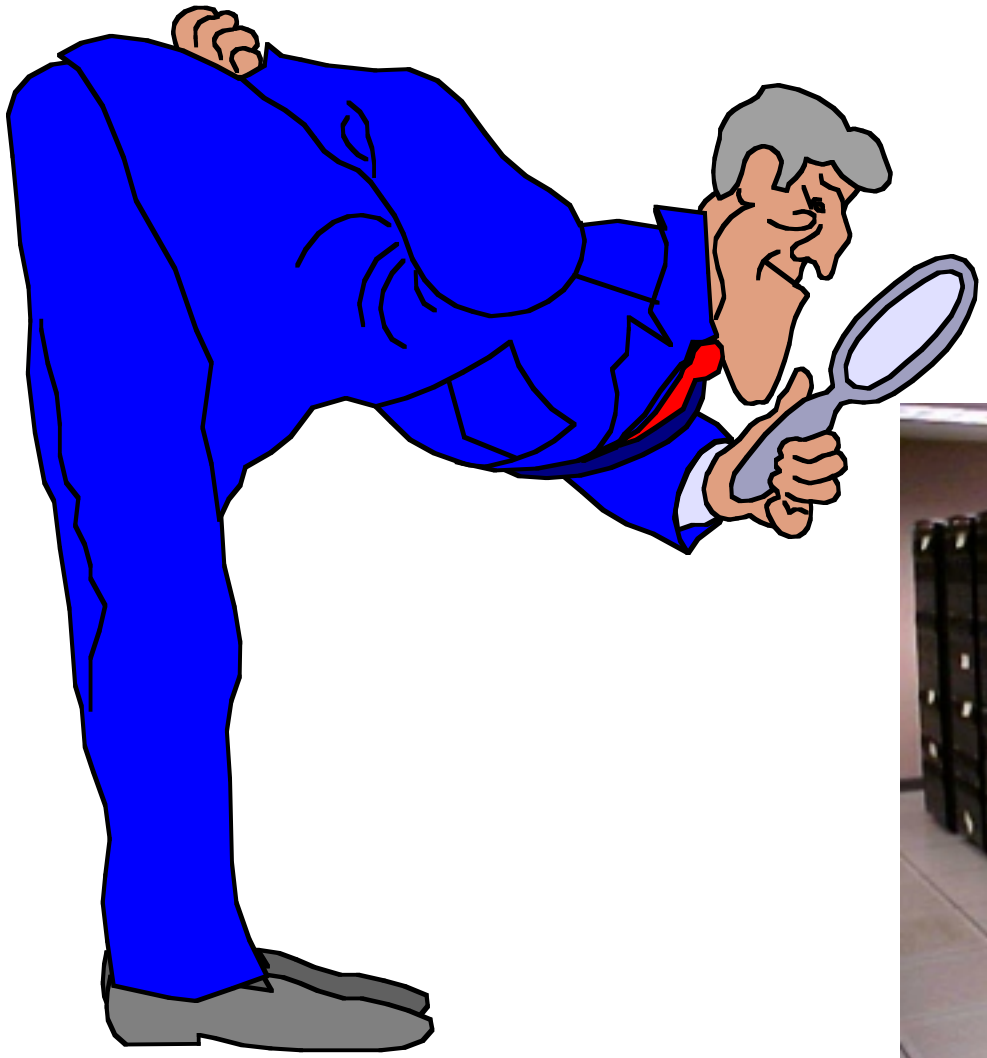


Total Storage Capacity 1.0TB

SAN Performance

AEROJET

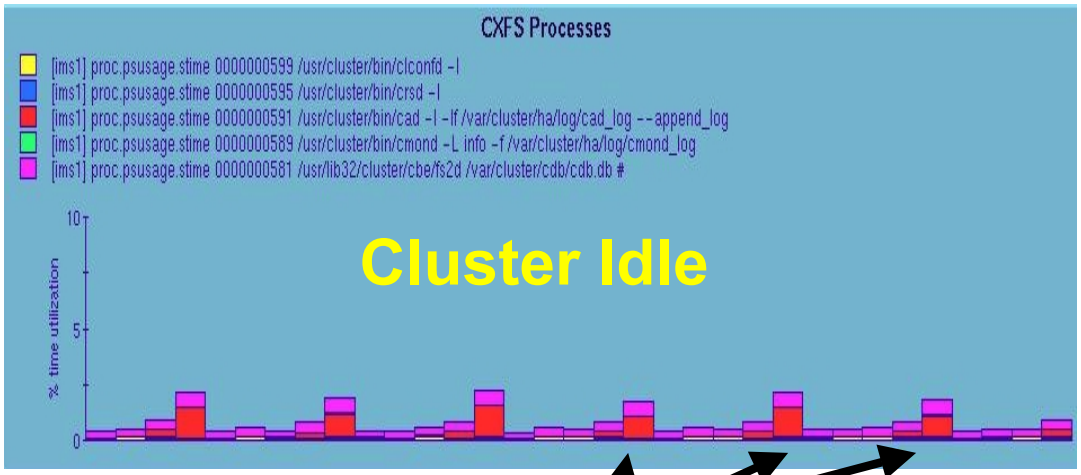
A GenCorp Company



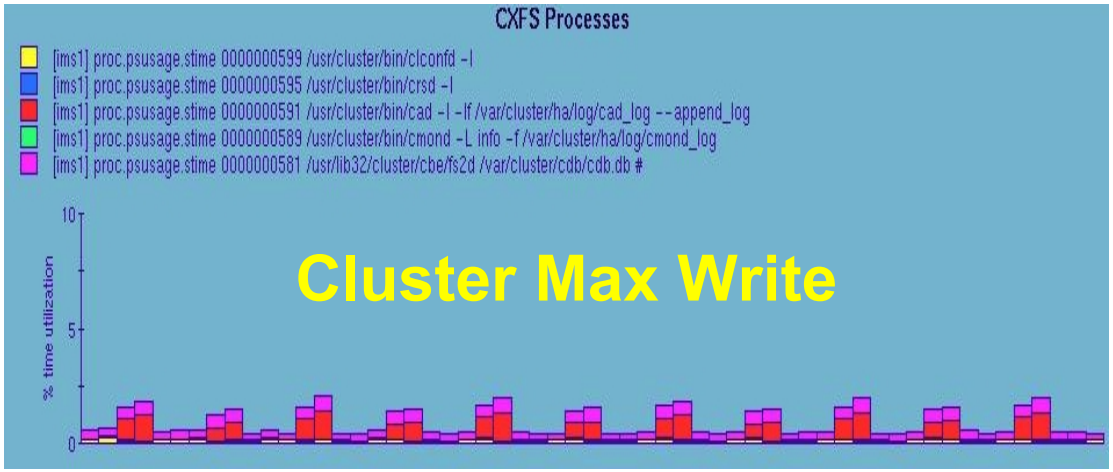
Processor Utilization

AEROJET

A GenCorp Company



Cluster Heartbeat



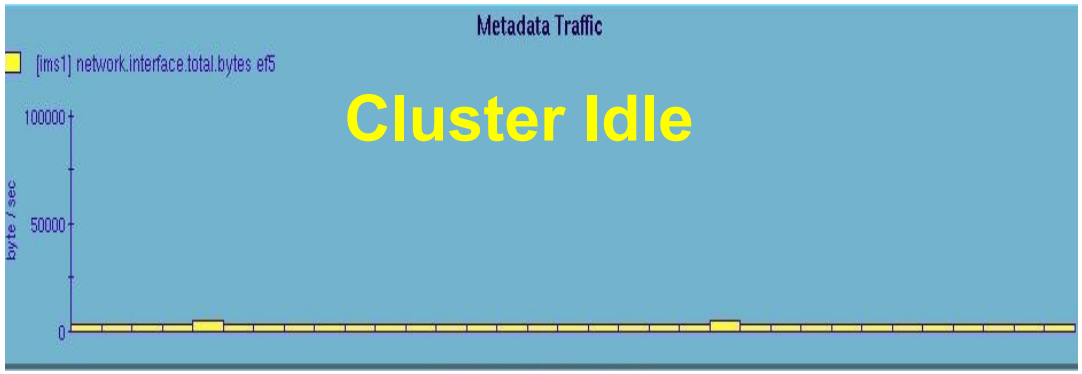
Test System:
8 Node Cluster Contains
2 – 32P Origin2000
5 – 16P Origin2000
1 – 4P Origin2000
Single CXFS Filesystem

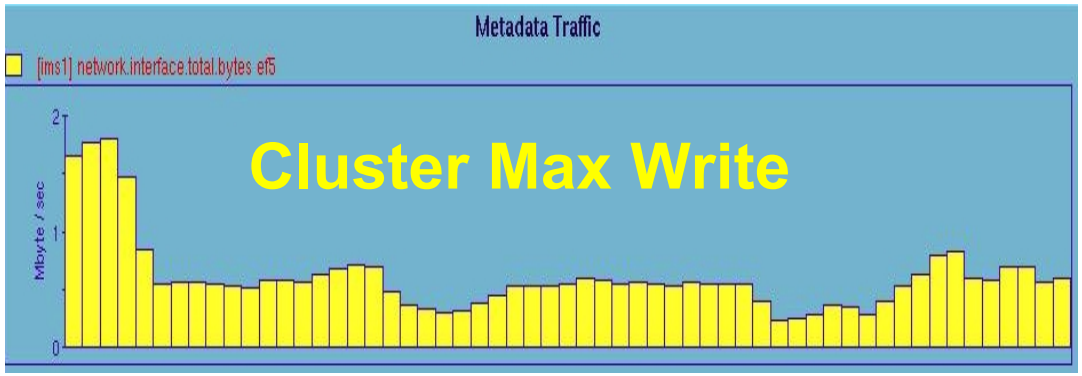
0.8 CPUs Utilized During Max Write on 8 Node Cluster

Metadata Network Performance

AEROJET

A GenCorp Company





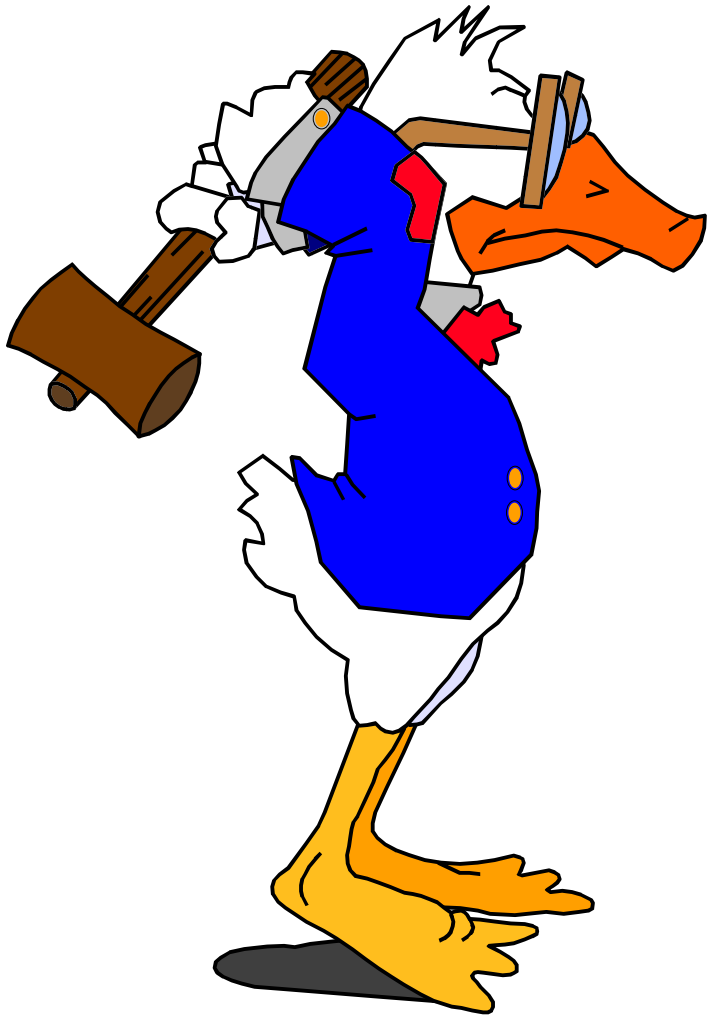
Test System:
8 Node Cluster Contains
2 – 32P Origin2000
5 – 16P Origin2000
1 – 4P Origin2000
Single CXFS Filesystem

8Mb/sec Ave During Max Write on 8 Node Cluster

Trials and Tribulations

AEROJET

A GenCorp Company



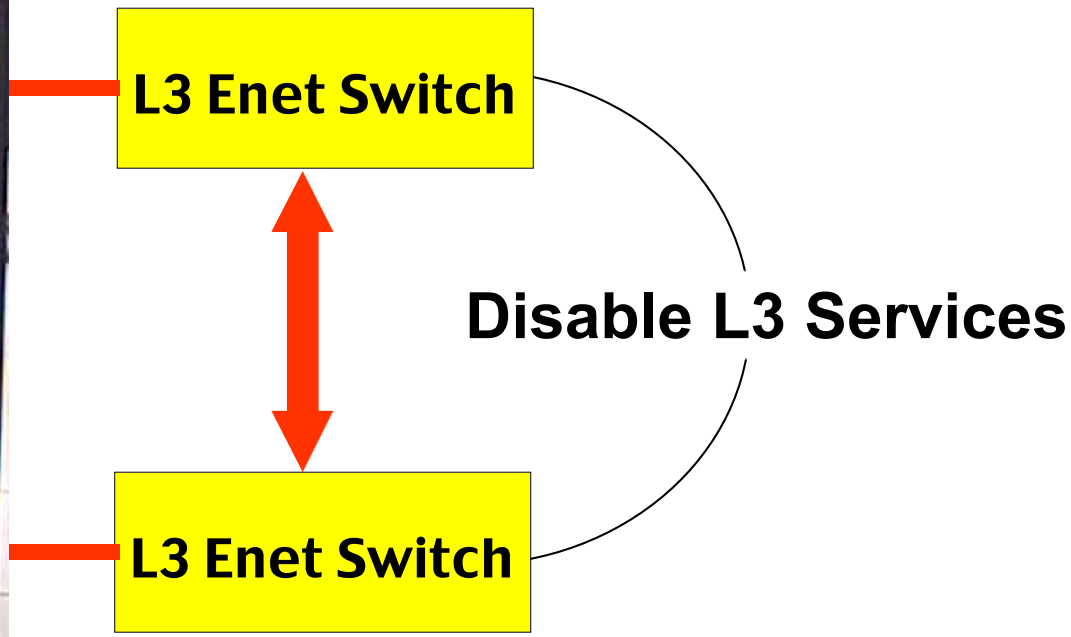
Layer 3 Networking

AEROJET

A GenCorp Company

CXFS uses Multicast protocol to transfer metadata

Does Not Conform to RFC 1112



If nodes are on different subnets CXFS will use TCP/IP

Two Node Cluster

AEROJET

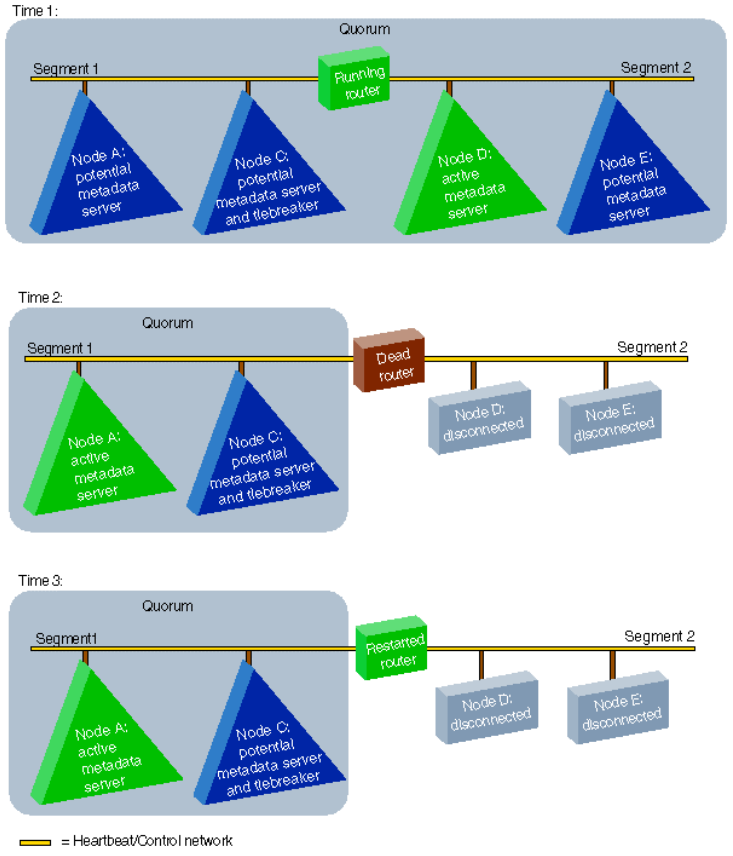
A GenCorp Company

HW Reset Cable

L3 Enet Switch

L3 Enet Switch

Add third node



One node failure can bring down both nodes

AEROJET

A GenCorp Company

Where's the Data?

Extreme detailed filesystem layout critical

WWN locations **Stripes & Slices**

Partitions **Local vs Cluster**

SAN Physical Wiring

AEROJET

A GenCorp Company

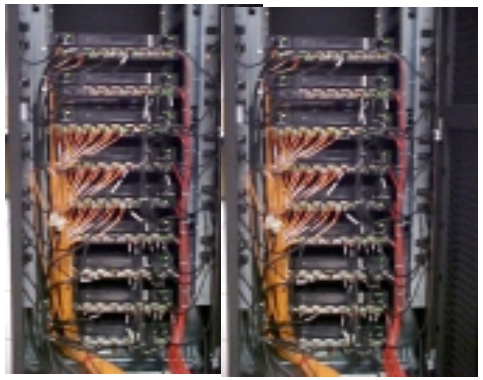
Fabric A Wiring

Fabric B Wiring

Installed System Layout

AEROJET

A GenCorp Company

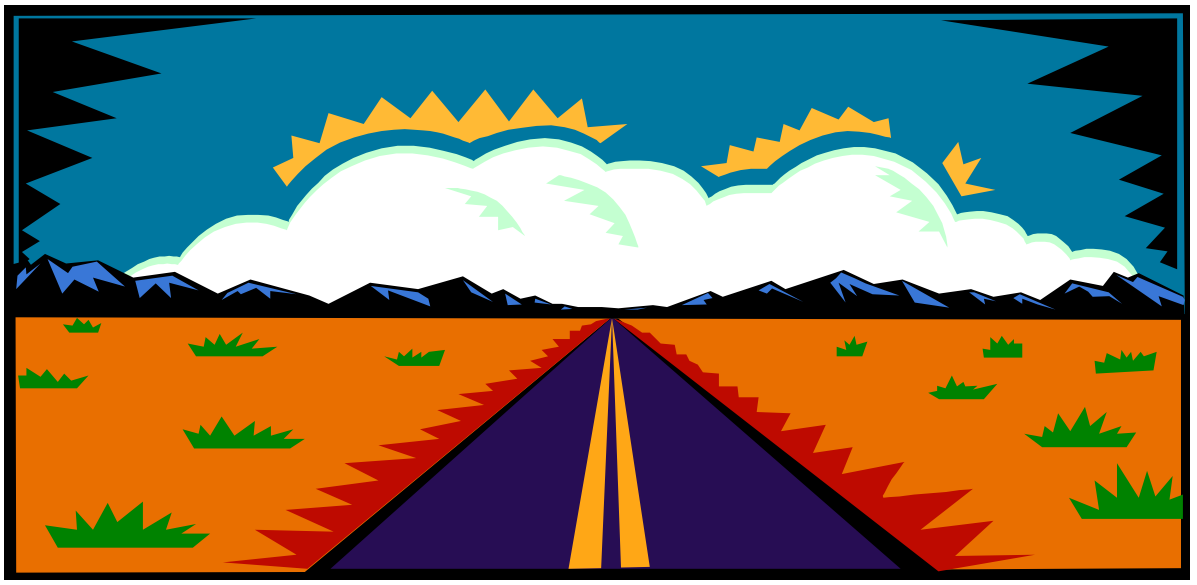


The Road Ahead

AEROJET

A GenCorp Company

- Larger SAN configurations are needed up from current supported configuration of 16 nodes to 64 nodes and beyond
- Heterogeneous SAN architecture using CXFS to support integration of other platforms
  - NT
  - Linux
  - Solaris

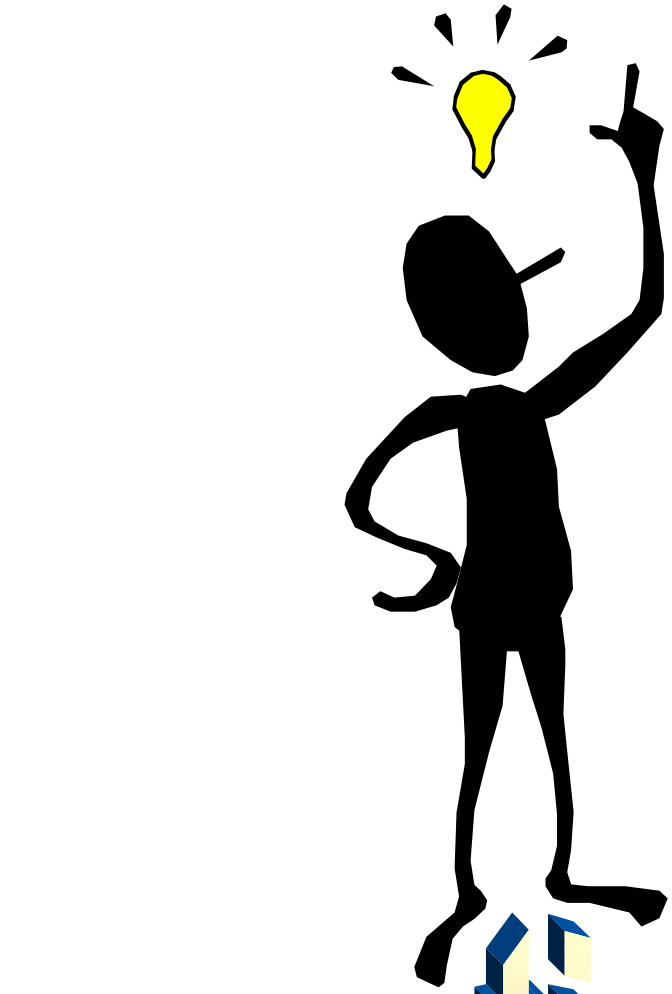


Summary

AEROJET

A GenCorp Company

- SAN performance exceeded our system requirements
- Low utilization of system resources
  - CPU
  - Network
- Close cooperation between Aerojet and SGI made this architecture a reality
- Carefully balanced risk vs reward made large-scale, highly diverse SAN integral part of real-time data processing network



Questions?