CREDIT SUISSE

IEEE 802 Plenary Session July 16-19 San Francisco

# Service Oriented Fabric: Datacenter Interconnect Requirements From Financial Industry's Perspective

Date: July 2007
Produced by: "Head" Bubba
IT  Research & Development
contact: Head.Bubba@credit-suisse.com
Head.Bubba@ieee.org

The materials may not be used or relied upon in any way.

# Service Oriented Fabric

## Convergence = Consolidation + Virtualization

- All traffic on one wire: LAN + FC/StAN + IPC/SAN =>
  - Virtually, but safely, separated traffic:
    - Trading
    - Storage (FCoE and RDMA / iWARP)
    - Customer login (secure access to internal network)
    - Development / R&D
    - Management
  - Quality of Service (QoS) Provisions
    - No Loss (Storage and IPC)
    - Delay guarantees for trading traffic (and customers)
    - Bandwidth guarantees for storage
  - Congestion Management (CM) and Load Balancing
    - Under application control (API)
  - Deadlock Management (DM)
    - DLKs are more catastrophic than congestion => must have solution
  - Destination (DST)-based per-flow RX rate calculation (throughput accounting )

# Service Oriented Fabric
## Key Needs For Datacenter Applications

- Must not have:
  - Traffic mix on same lane (different classes merged by switch or NIC)
    - Need to adhere to Service Level Agreements (SLAs)
  - Link-level loss (CM is no match for a fast LL-FC)
  - Mis-ordering
  - Deadlocks
  - Congestion
  - SlowStart (ramp-up a la TCP)

- Must Have:
  - Standards based solution for above not haves (no propriety solutions)
  - Clock Synchronization
    - IEEE 1588 protocol for high precision clock sync over Ethernet looks promising
  - Measurement -
    - Need to know what the latency and throughput are, and how they change over time
  - Lossless
  - Virtualization
  - HCAs & Fabrics must be able to interoperate with any other HCA

# Service Oriented Fabric

## Key Needs For Datacenter Applications

- Need to enable Virtual Data Center
  - Transform the data center to service provider from an equipment warehouse
  - Ability to create "application container" that can be moved from one server to another without service disruptions
  - Architecture handles problem not upper layers
    - Container must operate in hardware independent space

- The Entire Fabric Virtualized - Creation of a Virtual Resource Market
    - Ability to start application on any server and available fabric that meets SLA
    - Ability to live migrate an application to another server and/or fabric to meet SLA without having to know network topology
  - Clients, Servers, Interconnects, Routers and Switches
  - Defined by Objects
    - Polymorphism – allows the Open Service Oriented Fabric to interface with other Service Oriented Architecture Components
    - May be represented by XML

# Service Oriented Fabric

## Selected Pain Points

- Current link-level flow control (LL-FC) PAUSE causes 2 problems
    - Deadlocks
    - Saturation trees
  - Per-Prio PAUSE (PPP) emulates VLs/VCs => traffic separation
    - Does it have to remain proprietary in Ethernet?
  - Some applications are very sensitive to ordering...

- Deadlocks: Must be dealt with thru a standard solution.
  - Circular dependencies are possible: request/reply apps, greedy routing, buffer sharing...
  - VC-based solutions are known (ex: BlueGene, Cray etc.)

- Congestion: Must be dealt with thru a standard solution.
  - DCs need a stable, robust and efficient alternative to TCP
  - CM must be co-designed together w/ QoS (app-level SLA thru TX selection)
  - Need to easily manage Hotspots via SLAs

# Service Oriented Fabric

## Key Elements of Service Orientation

- Congestion needs to be handled at L2 not at Transport Layer
  - Needs to be addressed at L2 not at L4
    - Currently contradicting proposals?
  - Consider a Virtual Network: if you move a particular socket (along with application that uses it) from one machine to another, all of a sudden it must change its behavior because a network path has changed and congestion control of TCP must be adjusted accordingly

- Tools
  - Mechanism to provide latency measurement statistics
    - Need non-intrusive performance statistics that are available to check SLAs
      - Non-intrusive port mirroring
    - Need to know end-to-end roundtrip performance
    - Some have demonstrated a speedometer to show bandwidth performance
  - Measurement resolution to microsecond
    - Timing mechanism should introduce additional latency
    - If measurement is active device, then it should fully redundant

# Service Oriented Fabric

## Key Elements of Service Orientation

- Guaranteed SLAs For Bandwidth & Latency
  - Need exclusive bandwidth to meet application data & storage requirement SLAs Management - Not just fixed priority-based... Must be dynamic
  - Allow to dedicate guaranteed bandwidth for an application that no other application can use if the SLA requires it
  - The desire is to have well defined, deterministic latency
    - Currently, applications send data via TCP, and we've seen high variance in ack times

- Natively Lossless and Ordered Fabric
  - Some applications can NOT tolerate loss of data
  - No out-of-order delivery
    - Absolutely no re-ordering of transmitted data
  - Not just within a switch and datacenter (intra-DC), but also between datacenters and clients (inter-DC)
  - Lots of data now sent via IP multicast; while it may well from a latency standpoint, a lot of time is spent debugging packet loss / and ordering issues

# Service Oriented Fabric

## Key Elements of Service Orientation

- Quality of Service (QoS), Congestion Control and Adaptive Routing must be fully implemented
  - Needs to be “real” => standard products
  - No more talk of “we have enough bandwidth, so we don’t need it”
  - Do not want something (e.g. CM) to arbitrary slow down an application’s injections
    - Service Level Agreement (SLA) must be adhered to
  - A mechanism to pick-up possible future congestion must be in place that will send feedback to a SLA manager that will decide what should be done
    - Live application migration to another fabric is an option that should be considered
  - Clock Synchronization for all hosts on fabric

- Interoperability & Standards
  - No single vendor solution; at least two or three vendors will be used
  - Management, Virtualization and Protocol must be able to interoperate across different vendors
  - Can not have one fabric support a protocol that another fabric can not communicate with because it is not supported

# Service Oriented Fabric

## Need For A New Fabric

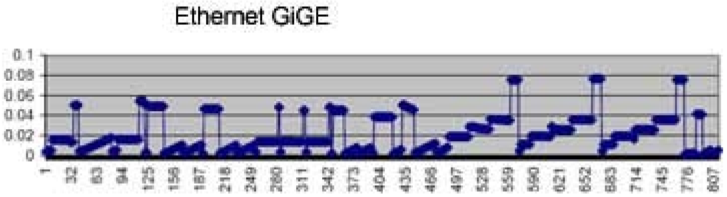
- One to many - being able to multicast out a stream of data to multiple endpoints efficiently
  - Current solutions include multi-cast NACK style protocols with the ability to separate streams (i.e. topic routing using multicast address)
- One to one - simple node to node reliable communication with minimal latency
  - This is both internal and external (LAN and WAN) and needs to include an efficient ACK like protocol and ordering algorithm.
  - Messages tend to be very small (less then one frame)
- The rate of market data traffic is always increasing with a significant impact on both the network and server infrastructure
  - it is projected for January 2008 there will be a need to be able handle ~6 billion messages per day for one market data feed *(estimates for one stream ~720,000 msg/sec, bandwidth including re-transmission ~450 Mpbs - 2 redundant streams available)*

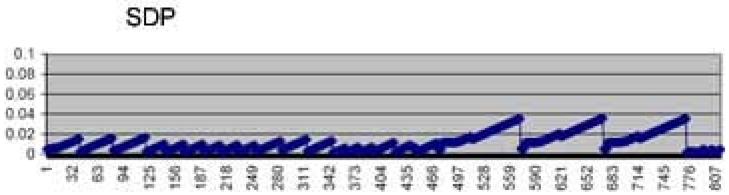
# Service Oriented Fabric

## Need for A New Fabric

- TCP/IP deficiencies
  - With TCP/IP, we have seen latency spikes (~40ms) caused by Nagle algorithm and the congestion avoidance window
    - Nagle disabled => still latency spikes
      - when connection is idle for 5 minutes (unless continuously keep setting QUICKACK is not a viable solution)
  - SLOWSTART is unacceptable for some classes of applications
  - TCP/IP offloading & iWARP at this point time doesn’t seem to address this problem
    - At most it may mask/reduce it

- A better solution is needed for our applications
  - “Jitter” could equate to a measurable loss of revenue
    - Testing has demonstrated predictable repeatable “jitter”
  - Need consistent low latency predictable message delivery

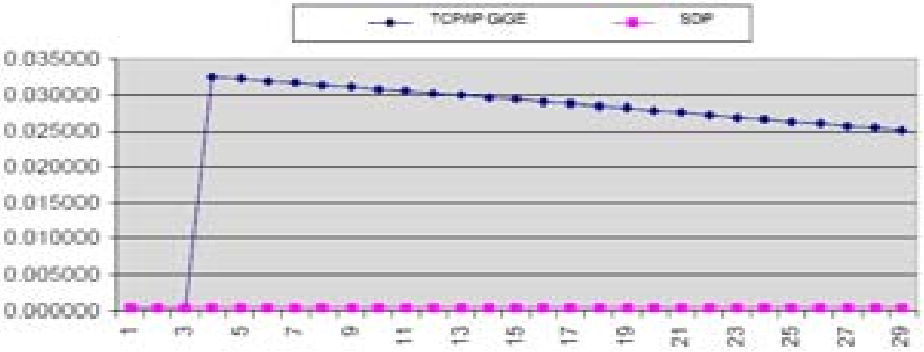

Ethernet GiGE


SDP

# Service Oriented Fabric

## Need For A New Fabric

- What exists today is not good enough for tomorrow – we need something better
- TCP/IP Client Latency (GiGE) as compared to SDP on another fabric
  - Ethernet needs to be more predictable & have better performance
  - more un-managed bandwidth is not the solution
    - wider roads invite more traffic... and ensuing congestion
  - The way things are done today does not make them right for tomorrow



# Service Oriented Fabric

## Need For A New Fabric

- TCP/IP is 20-yrs. old => many major apps are socket-based...
  - Enterprise customers can not re-write millions of lines of code
  - Need deterministic behavior
  - SDP is a good start

- Need Low Latency / Need Near Fabric Performance
  - We need an off-the-shelf easy plug-in to get instant Return on Investment (ROI) or the High Speed Interconnect won't get wide deployment
  - Message sizes 2bytes - 60bytes, 0.5k – 2k, 8k at most?
    - ***Need to optimize for small message sizes as well not just large!***
  - Can not arbitrary decide to slow down an application sending data
  - Need Multicast solution

- Some applications can not afford to lose data -This could equate to missing an opportunity to make an automated decision that could result in the loss of $$$

# Service Oriented Fabric

## Governance and Validation

- Testing and Validation needs to part of any roadmap deliverable – code without validation should not be released
  - Software needs to have a certification to pass
  - Can not have pick and chose the features you want to support
  - Which leads to…

- Software needs to be tested vigorously on all supported platforms
  - End users, ISVs and OEMs should be a part of an interoperability lab

- Only one supported snapshot
  - Do not want Company A, B … Z  to take a snapshot in time, and have their own internal staffs supporting their snapshot version
  - We want companies to support the same snapshot, and all work together on bug fixes & enhancements

# Service Oriented Fabric

## Value Proposition

- Co-locating client / server tiers for demanding trading applications
  - Best of breed, ultra-performing network, server and storage technology to support demanding transaction workloads
  - Unprecedented performance of core system interconnects
  - Unbound transactional capabilities and superior resource utilization in the OS

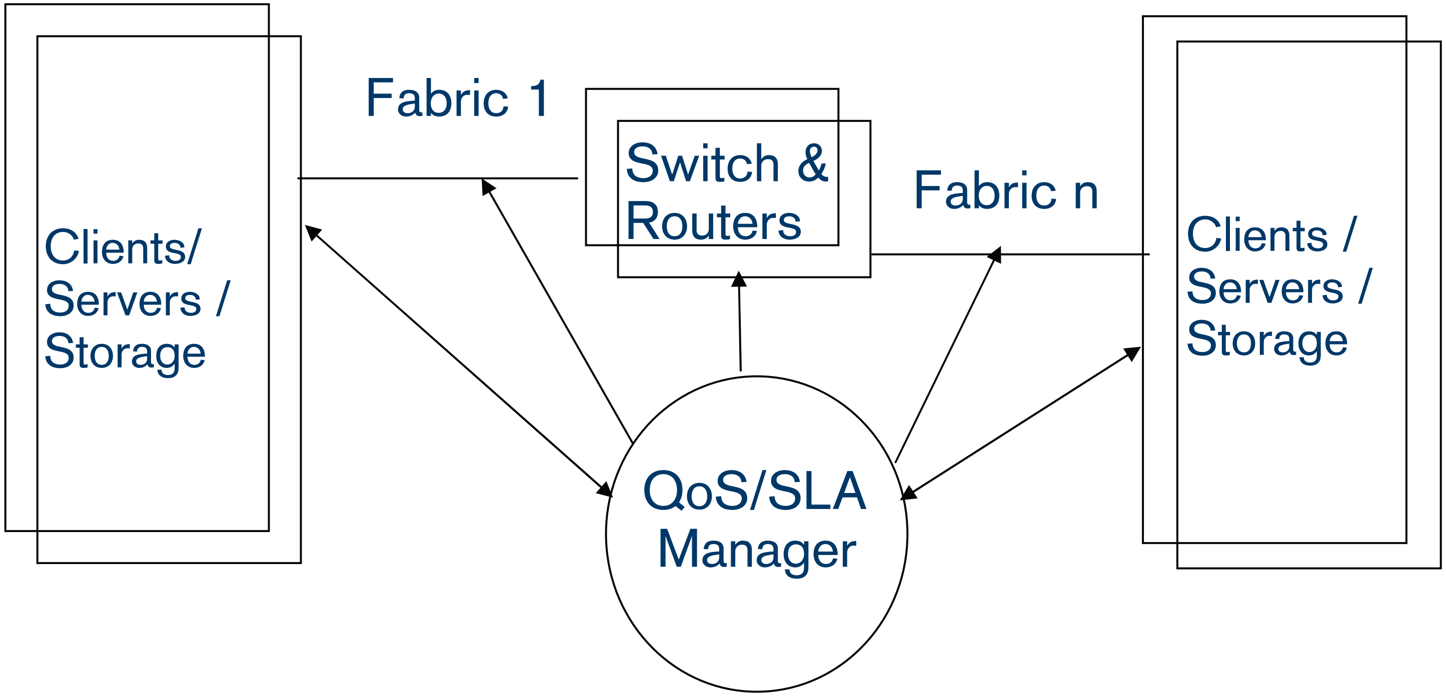
- Predictable application response times for time-sensitive workloads
  - This includes use of Real Time OS

- Enterprise GRID solution and calculation accelerators for quantitative business applications

- Open Solutions & Standards
  - Reduce complexity

# Service Oriented Fabric

## Example



# Service Oriented Fabric

## Q & A

- Contact Information
  - “Head” Bubba

    Credit Suisse  
    IT Research & Development  
    11 Madison Avenue  
    New York, New York 10010  
    email: head.bubba@credit-suisse.com  
    head.bubba@ieee.org