citihub™

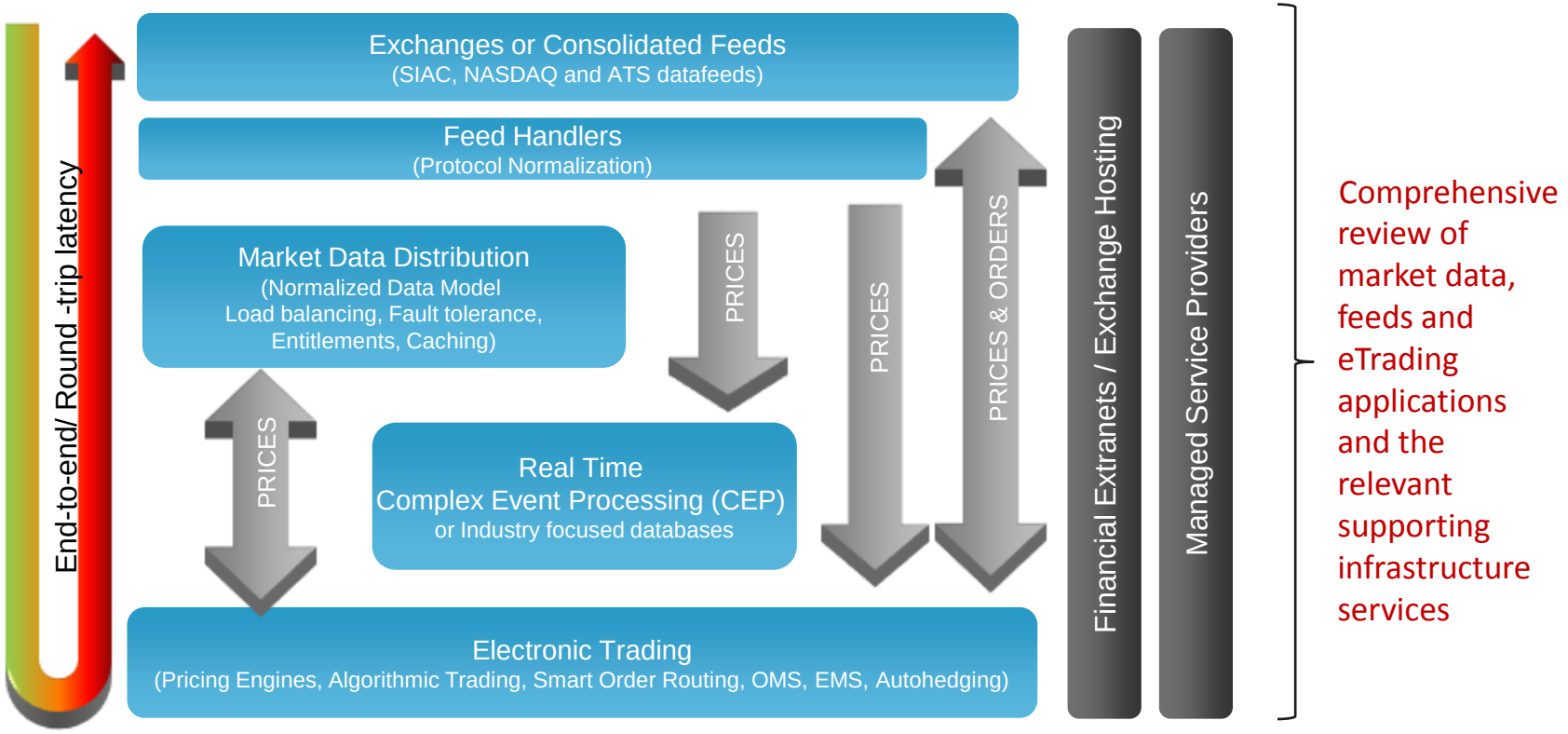
▸ eTrading & Market Data ▸ Agile infrastructure ▸ Telecoms ▸ Data Center ▸ Grid

A Holistic Approach to Eliminating Latency

April 2009

Smart Infrastructure Solutions
London • New York • Singapore
www.citihub.com

Citihub Latency Framework



Comprehensive review of market data, feeds and eTrading applications and the relevant supporting infrastructure services

Interviews and observations

- Architecture analysis
- Technology maturity and fit-for-purpose
- Best practices and business usage
- Network and multicast/ messaging implementation

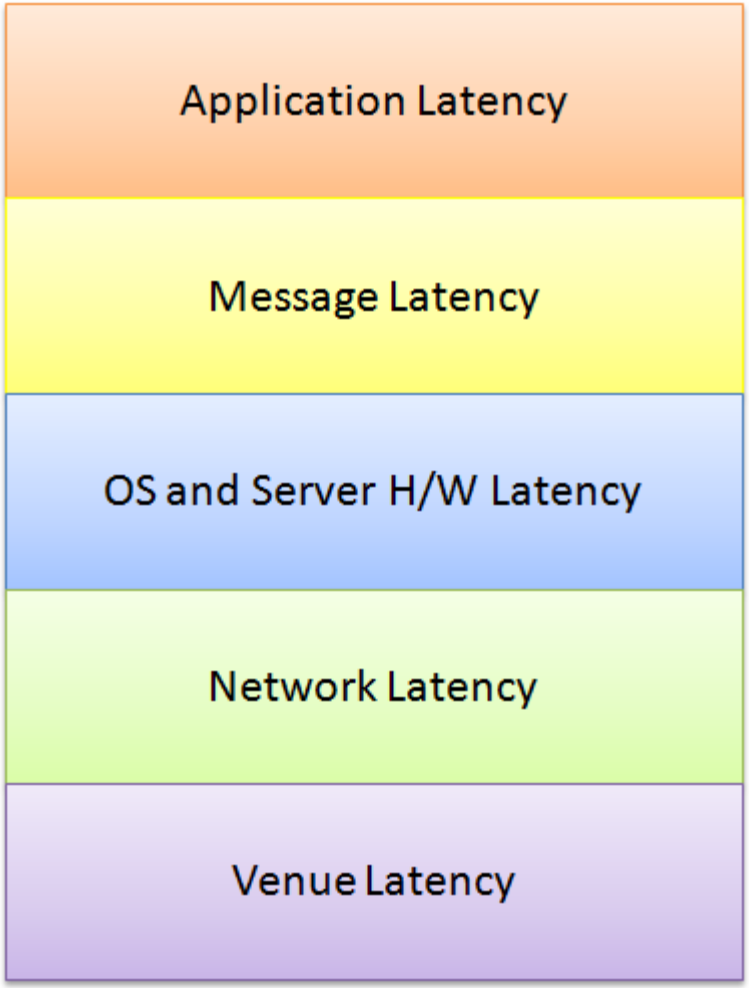
End-to-end review

- Central co-ordination, process, tools & risk reporting
- Full or 'Lite' assessment

Data and trend analysis

- Latency and performance metrics with distribution analysis
- Configuration tuned for low latency requirements
- Monitoring tools review
- SLAs and outage reviews

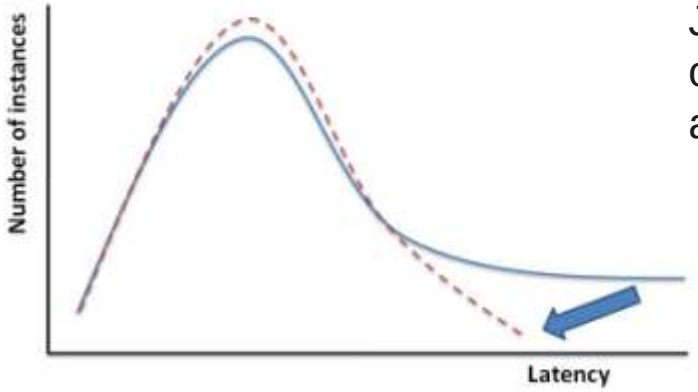
Latency Stack



What is latency optimization?

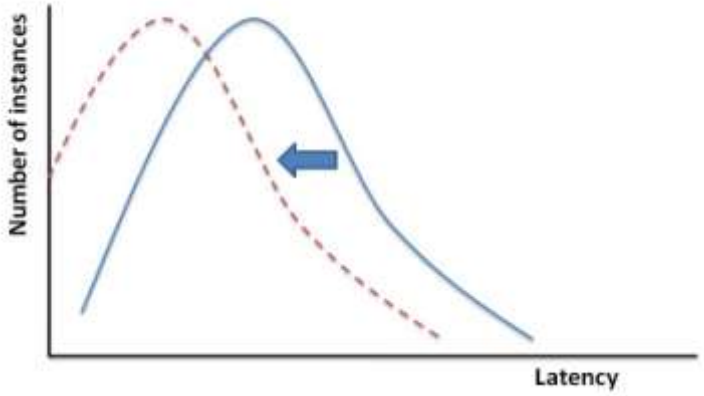
citihub

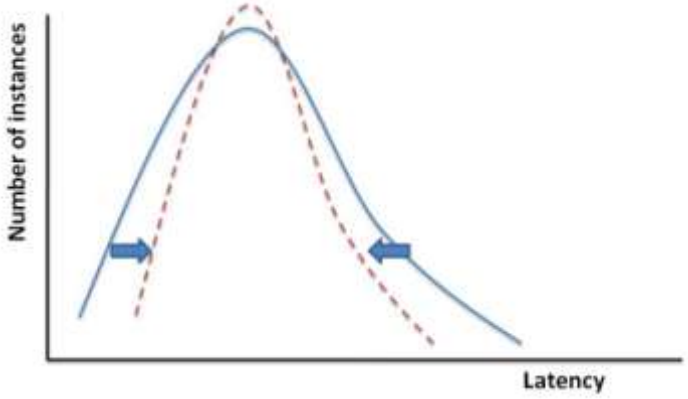
- Urgent focus to remove **event-driven latency** – removing causes of jitter to ensure huge volume spikes don't impact latency.



Jitter is a result of capacity issues, network and operating system packet loss, network congestion and application garbage collection.

- Once event-driven latency has been fenced, focus on improvements on **operational latency** – general reduction in mean latency





Optimising or using new technologies to reduce overall latency or make more deterministic

Remediating latency: areas of focus

citihub

- Review OS network statistics e.g. nocanputs, Drops, tx-queue to determine OS efficiency and tune OS
- Review router/switch network statistics for problem indicators e.g. Discards, Catalyst CPU utilization and link utilization
- Use Real-time OS distributions e.g. Redhat MRG, SuSE RT, Solaris to improve pre-emption and reduce jitter.
- Process binding and Interrupt fencing
- Long haul carrier selection and network optimization *
- TCP/IP bypass or Offload *
- Manage garbage collection for Java and .NET e.g. RT JVM *
- Replace Enterprise Firewalls with lightweight, role specific devices with simple rulesets
- Review Programming models and best practices, messaging and middleware choices
- Replace or front RDBMS with in-memory databases
- Deploy latency optimized appliances
- Exotic hardware platforms i.e. Nvidia CUDA, FPGA , IBM CELL

Latency monitoring

Real-time monitoring, forensics and event reconstruction, capacity planning, tuning and business reporting.

- Network discards reporting
- Application timestamping
- Microburst detection and network monitoring
  - Corvil
  - Correlix
  - NetScout
- Market data protocol decode and correlation
  - TS Associates Tip-off
  - Trading Metrics
  - SeaNet
- FIX protocol decode and correlation
  - Clearsight Networks Cronos
  - NetQoS Trade Monitor
  - SeaNet
- Data capture hardware (with Absolute time synchronization)
  - Endace Ninja
  - Niksun
- Software based algorithms to compare direct feeds
  - 4th Story

Telecommunications carrier sourcing – Europe example

European trading centres

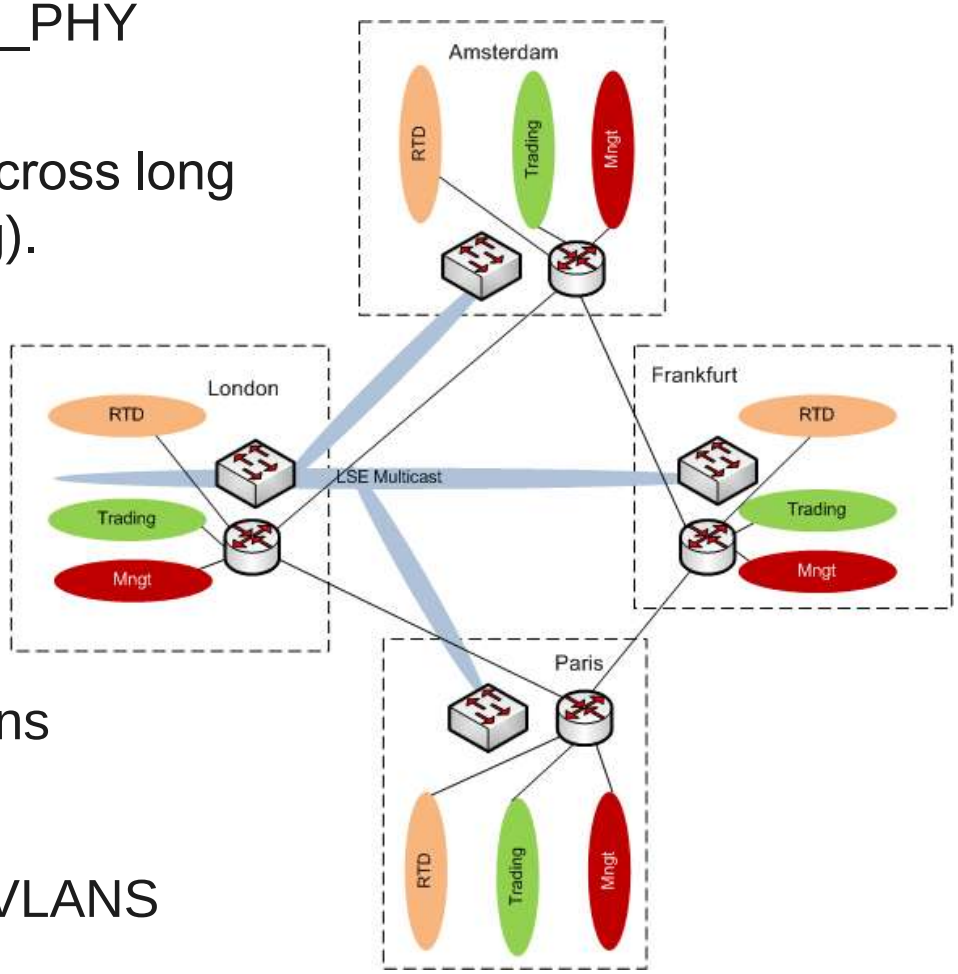
Route	Km	Theoretical Latency (RTT)
Amsterdam-Frankfurt	444	5.3 mS
Amsterdam-Paris	508	6.0 mS
Amsterdam-London	537	6.3 mS
London-Paris	455	5.5 mS
London-Frankfurt	769	8.6 mS
Paris-Frankfurt	572	6.6 mS

Theoretical lowest achievable latency

- Business requirement for multiple venue across a region or globally
  - Compliance, asset dependencies and arbitrage opportunities
- Carrier prices have never been better
  - 10Gb London – Frankfurt only €15k pm
- Significant differences in latency between carriers
  - Route dependant
  - Hops
  - Bandwidth availability
- Technology used by carrier makes a difference, especially during periods of high traffic bursts
  - DWDM preferred over MPLS
- Consider high performance WAN technology like stretched VLANS

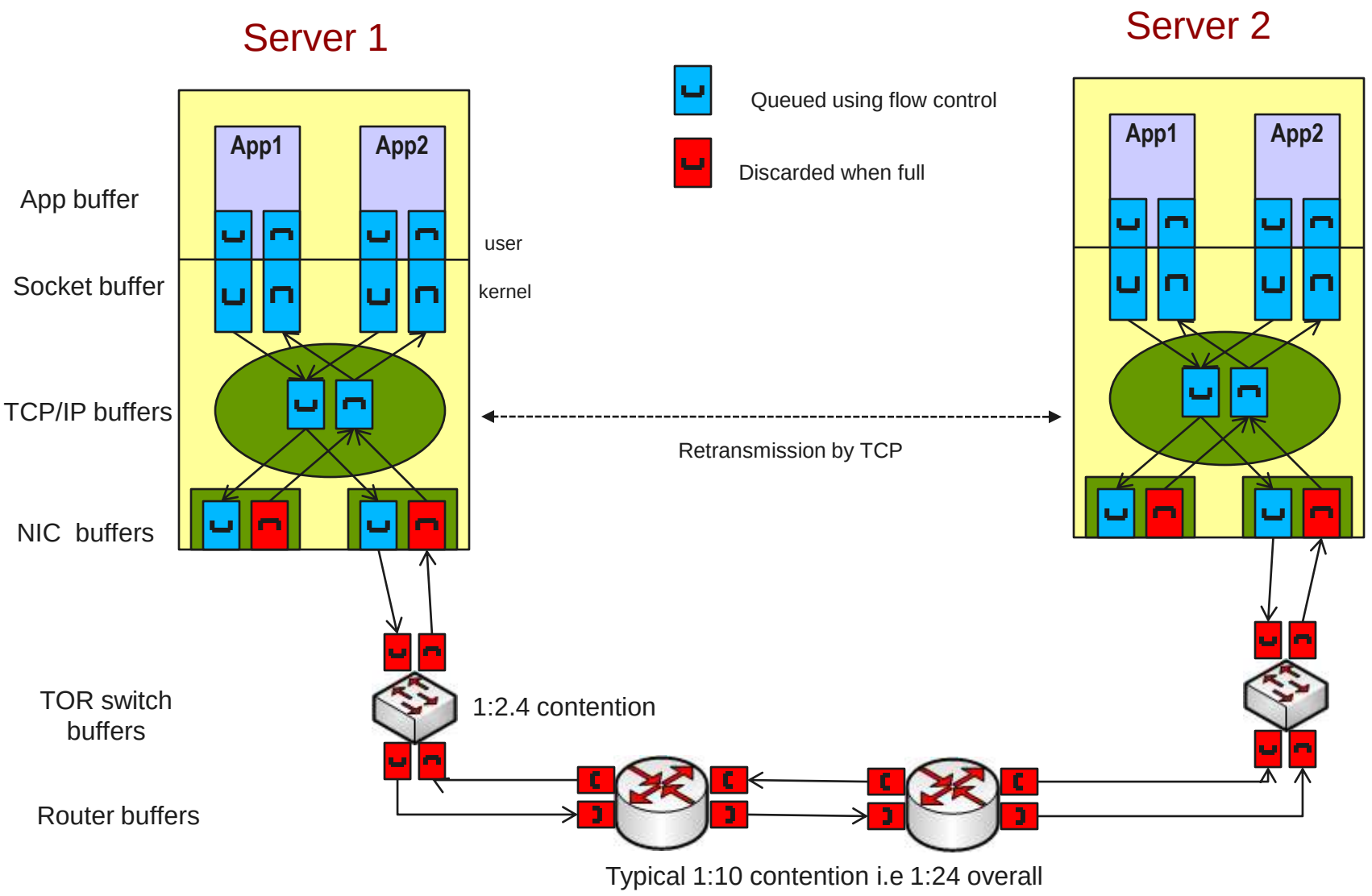
Stretched VLAN

- Low cost, DWDM or Ethernet LAN_PHY ports e.g. VPLS
- Selected VLANS are “stretched” across long haul circuits. (Actually L2 Bridging).
- Creates a single subnet
- Can be used for:
  - Raw multicast feeds
  - DR clustering
  - VM migration
- Dependent on spanning tree options
  - Some switches are better than others
- Share physical circuit with routed VLANS
- Traffic shaping to manage bursts

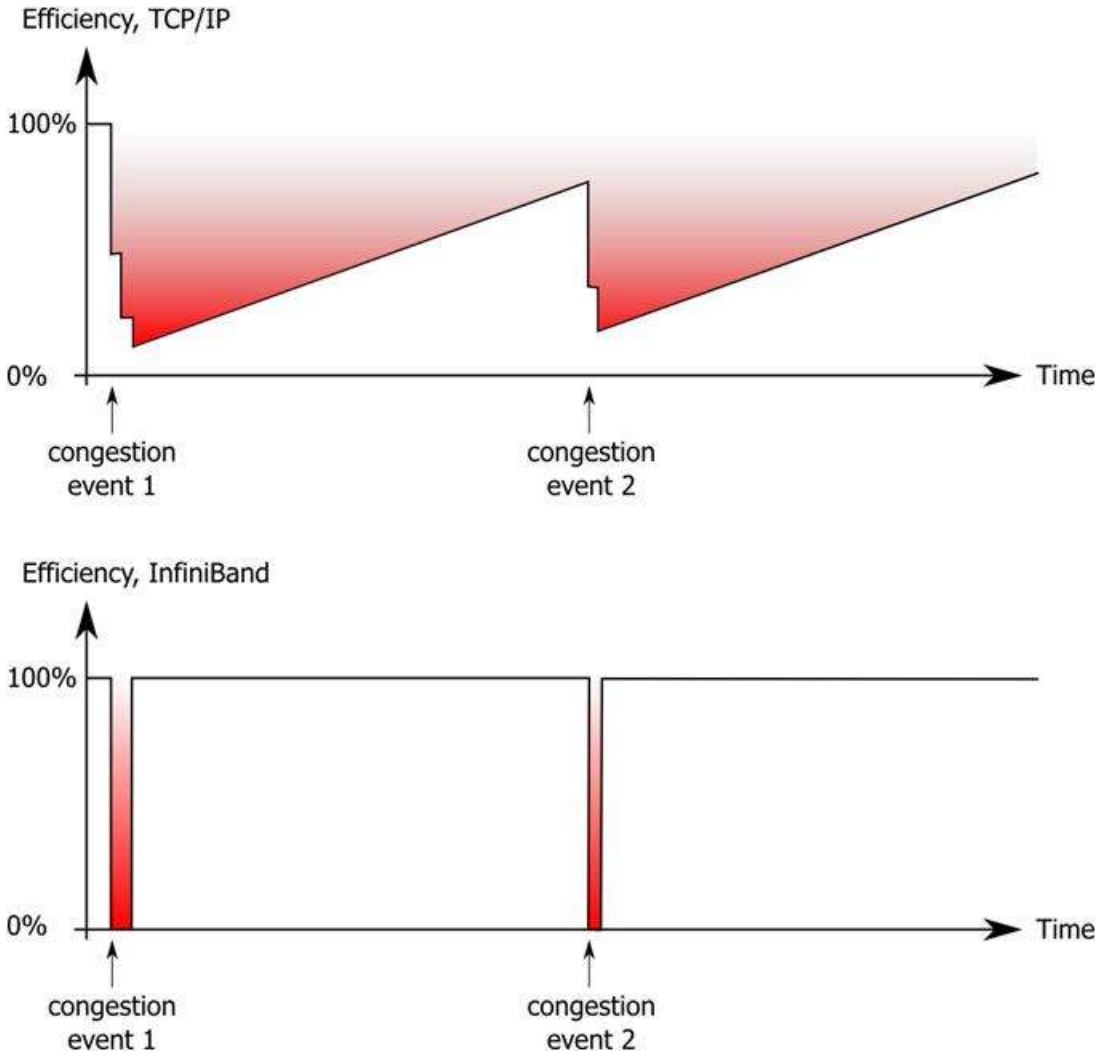


Ethernet Packet Queues, Drops and Discards

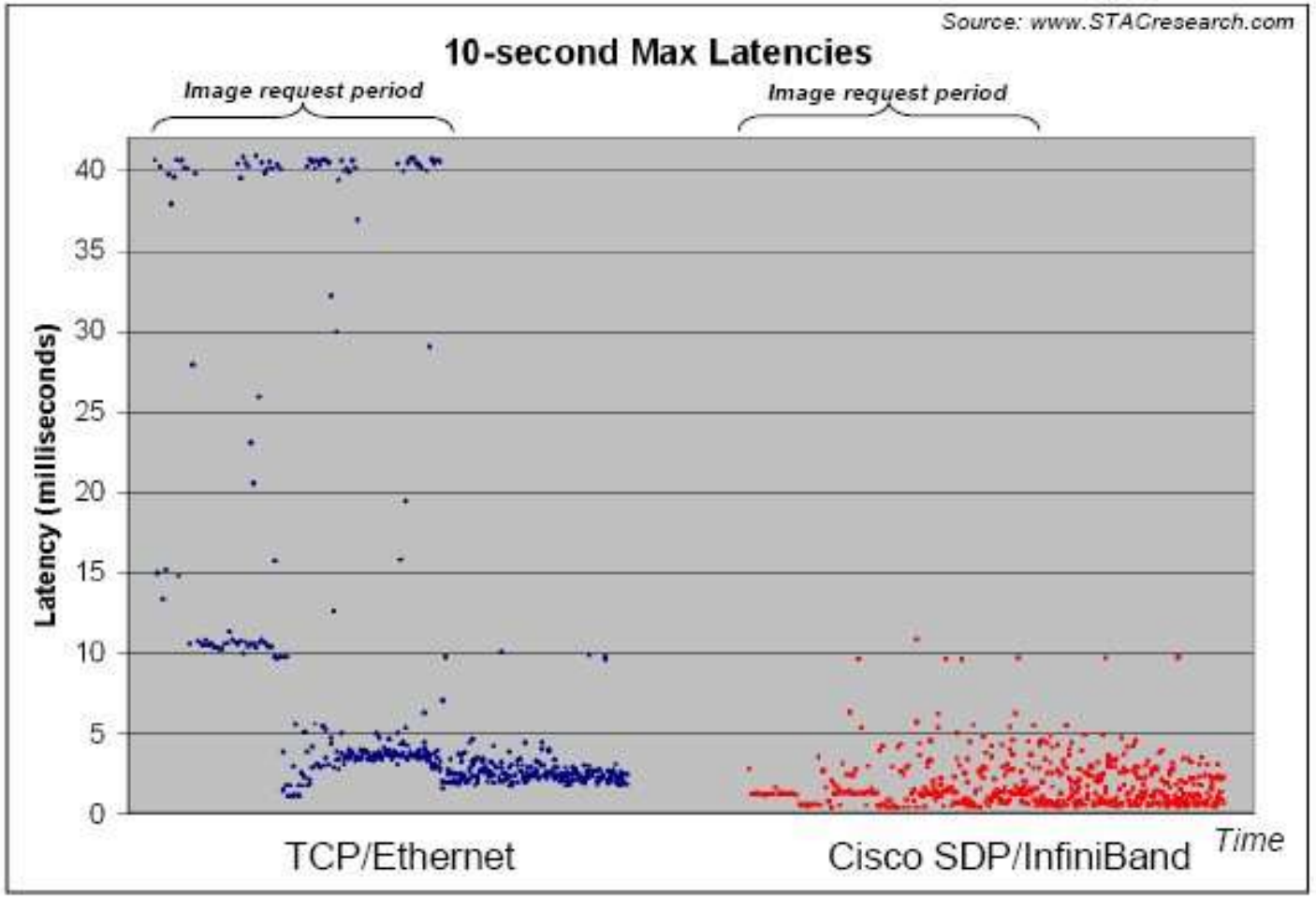
citihub



TCP/IP recovery for congestion events



TCP/IP/Ethernet impact on latency and jitter



TCP/IP bypass

- Options include TCP Offload engines or protocol bypass
- OpenFabric Alliance delivers OFED stack for Linux and Windows. Option for all major Linux distro's. Default for RH MRG and SuSE RT distro's
  - Supports both InfiniBand and 10G Ethernet with iWARP
  - Only accepted RDMA implementation for Linux kernel
- API's include:
  - Sockets Direct Protocol (SDP)
  - Reliable Datagram Socket (RDS)
  - Direct Access Programming Library (DAPL)
  - NFS/RDMA
  - SCSI RDMA Protocol (SRP)
  - iSCSI over RDMA (iSER)
  - API's are strategic. Decide tactically whether to deploy InfiniBand or 10GE
  - Can hedge further by using dual mode adapters e.g. Mellanox ConnectX VPI

Multicast

- Defined in InfiniBand Architecture (IBA), sends multicast using unreliable datagrams. Switches replicate packets as required.
- IETF define multicast mapping for IPoIB, which has been implemented in OFED
  - "standard" Ethernet multicast programs just work
- High performance achieved by programming at the VERB level
  - Voltaire Messaging Accelerator (VMA)
    - 1Ge = 47µS, 20G InfiniBand/IPoIB = 21.5µS, 20G InfiniBand/VMA = 4.4µS
  - Cisco Datagram Acceleration Layer (DAL) Protocol
  - Need an "open" implementation - potential projects to derive from:
    - Cern LHC Detector
    - Ohio MPI-bcast

Java low latency?

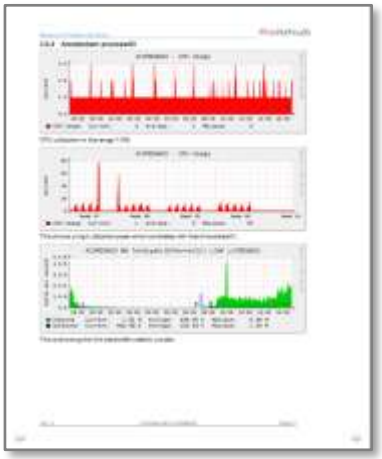
- Garbage Collection stalls causing jitter
  - JRE 1.5 use concurrent low pause GC
  - Move to Real Time JVMs or Azul
    - BEA Weblogic Real Time
    - Sun Real Time Java System
    - IBM WebSphere
  - RTSJ added real-time Threads, Scheduling, and Synchronization
  - Immortal memory – never destroyed
  - Scoped memory – can be GC when program is outside of scope
  - Develop GC “enlightened” Java code
- Serialisation
  - Distributed object models, e.g. RMI have to serialise the object graph dependencies and translate into byte types. This serialisation dominates the latency when high speed LANs are used

Java Fast Sockets – Open Source

citihub

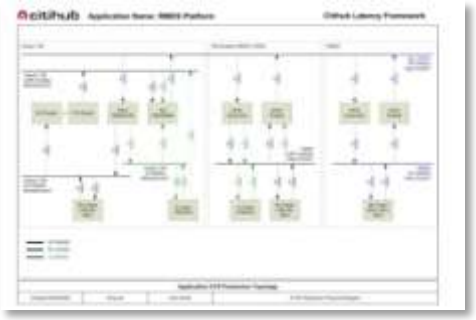
- Sacrifice portability by bypassing serialization.
- Extends socket API
- Implementations leverage RDMA based interconnects by allocating from discrete memory regions
- More an interesting research project than viable implementation but worth looking at where Java has to be used
- Also a problem for the HPC world – some crossover with Java-MPI activities

http://jfs.des.udc.es

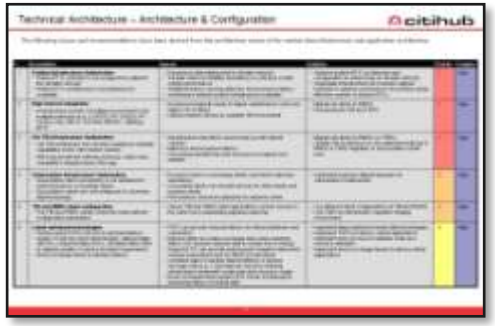


Performance metrics

Logical flow and bottlenecks



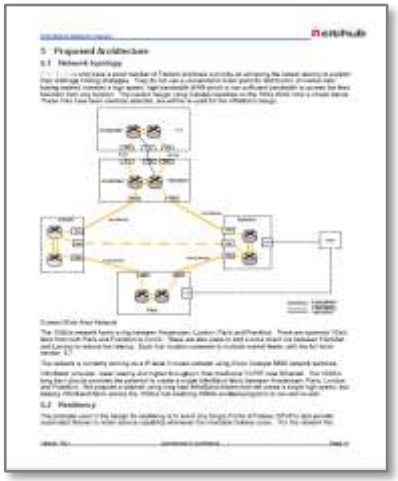
Technical documentation



Risks & impact

Latency report

Remediation plan



Ultra low latency reference design

Smart Infrastructure Solutions
London • New York • Singapore
www.citihub.com

Richard.Crouche@citihub.com