Abstract

Specialized hardware is giving architects new high-efficiency options to accelerate the WAN and avoid "long fat network" problems. This session will explore how network processors, FPGAs, flash storage, ultra capacitors, and other exotic silicon is increasing the capabilities and performance of WAN-based applications. Specific use cases include Distributed Message Routing, Web Data Streaming, Sensor Nets, and Active/Active Data Grid Replication.

Distributed Data Fabrics and Hardware WAN Optimization

Achieving 10X WAN Efficiency in Globally Distributed Applications

Hans Jespersen

Systems Engineer

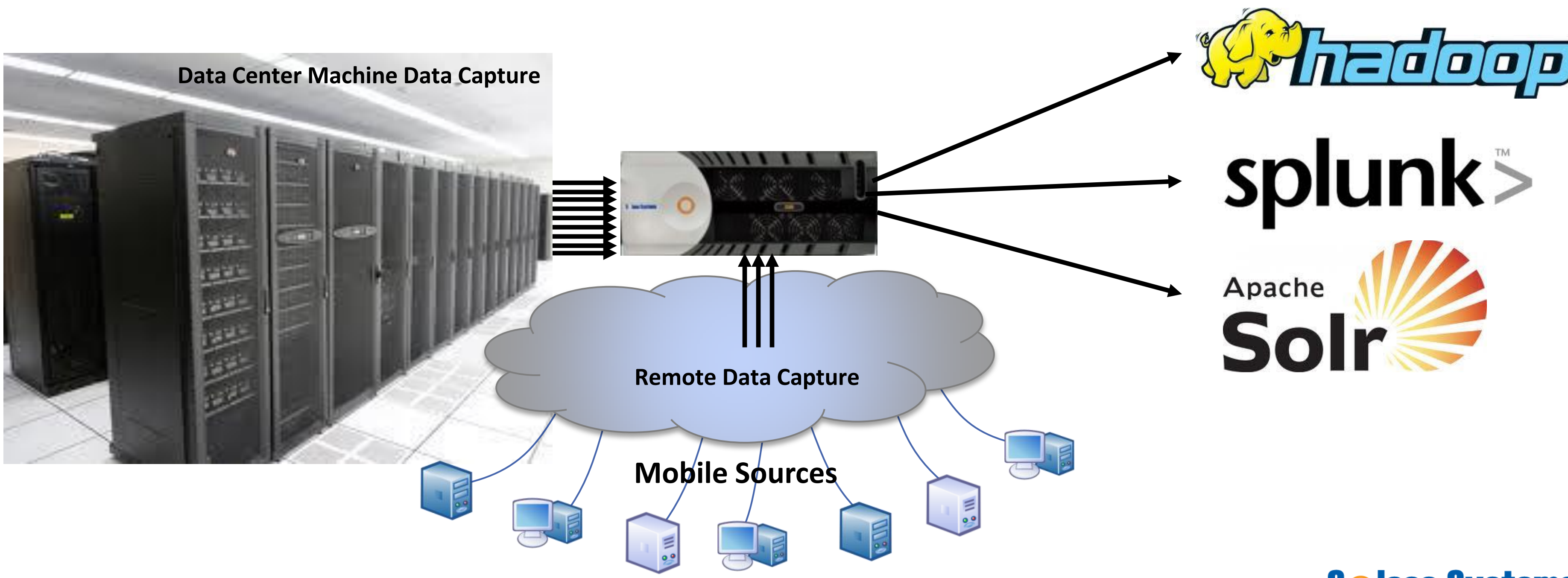
hans.jespersen@solacesystems.com

Agenda

- Introduce the use case
- TCP/IP and the Long Fat Network Problem
- Technology & Industry Trends
- How do traditional WANop solutions help (HW & SW)
- What isn't addressed with network layer WANop
- Message Brokers and application specific WANop
- Advanced Silicon and Exotic Hardware
- Benchmarking Performance
- Q&A

Real-time Streaming Big Data

*Need is for efficiently collecting, aggregating and moving large amounts of streaming machine generated data from **multiple sources** to **multiple data stores** across **multiple locations**.*



Old Faithful

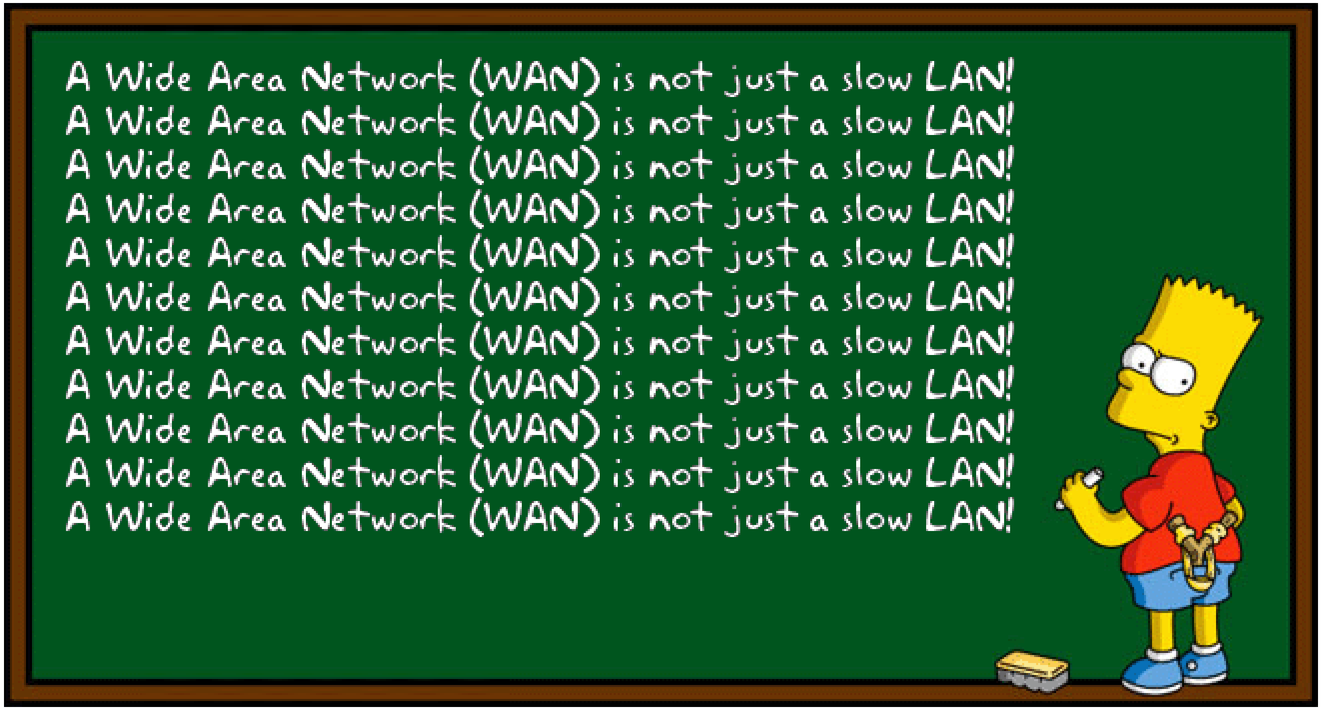
TCP Header

<table>
<tr><th colspan="2">Offsets Octet</th><th colspan="8">0</th><th colspan="8">1</th><th colspan="8">2</th><th colspan="8">3</th></tr>
<tr><th>Octet</th><th>Bit</th><th>0</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th><th>16</th><th>17</th><th>18</th><th>19</th><th>20</th><th>21</th><th>22</th><th>23</th><th>24</th><th>25</th><th>26</th><th>27</th><th>28</th><th>29</th><th>30</th><th>31</th></tr>
<tr><th>0</th><th>0</th><td colspan="16">Source port</td><td colspan="16">Destination port</td></tr>
<tr><th>4</th><th>32</th><td colspan="32">Sequence number</td></tr>
<tr><th>8</th><th>64</th><td colspan="32">Acknowledgment number (if ACK set)</td></tr>
<tr><th>12</th><th>96</th><td colspan="4">Data offset</td><td colspan="3">Reserved
0 0 0</td><td>N
S</td><td>C
W
R</td><td>E
C
E</td><td>U
R
G</td><td>A
C
K</td><td>P
S
H</td><td>R
S
T</td><td>S
Y
N</td><td>F
I
N</td><td colspan="16">Window Size</td></tr>
<tr><th>16</th><th>128</th><td colspan="16">Checksum</td><td colspan="16">Urgent pointer (if URG set)</td></tr>
<tr><th>20
...</th><th>160
...</th><td colspan="32">Options (if Data Offset > 5, padded at the end with "0" bytes if necessary)
...</td></tr>
</table>

What do we get?

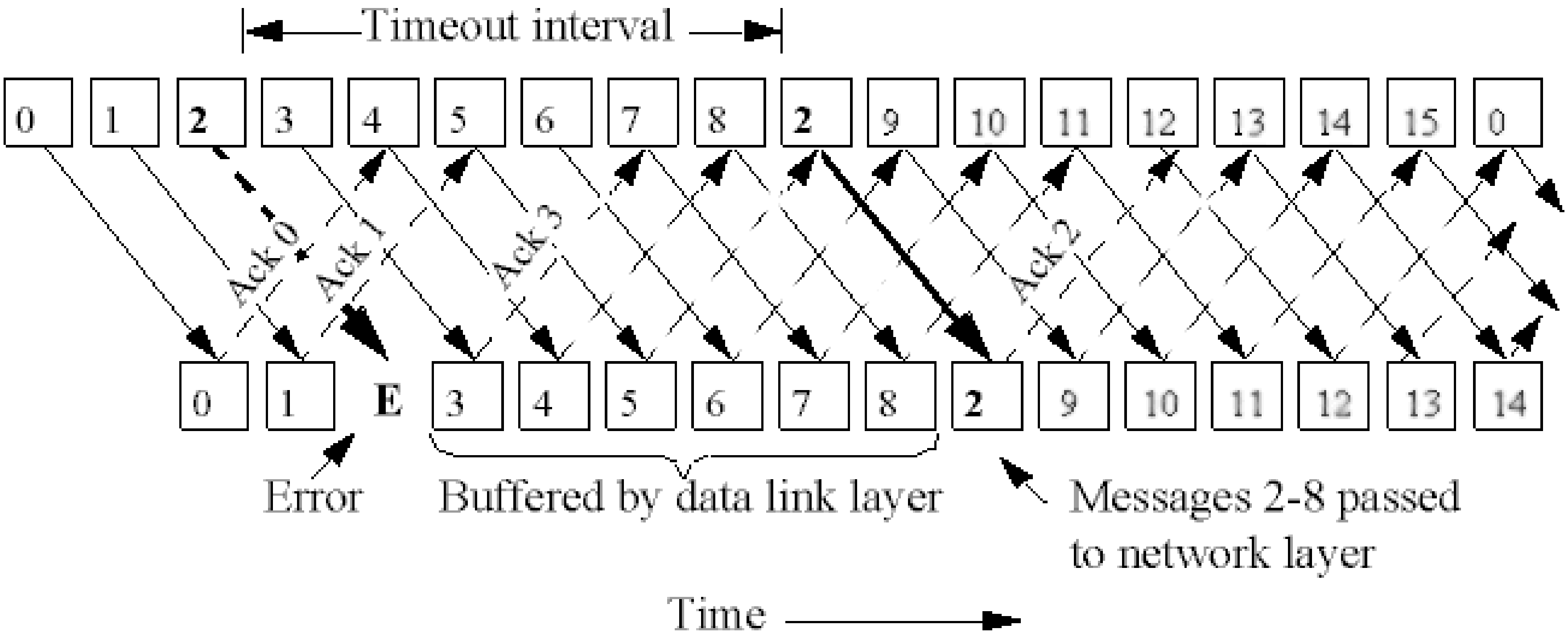
- Reliable
- Ordered delivery (of a stream of octets)
- Error-free data transfer
- Flow control
- Congestion control

Everything comes with a price



Solace Systems™

The LFN (Elephant) in the room



Throughput != Bandwidth

- Bandwidth
- Latency (RTT)
- Error Rate (Loss)

- Handy online calculators of effective throughput
  - http://www.silver-peak.com/calculator/

CALCULATOR

WAN Throughput Calculator

This calculator is used to measure the effect that bandwidth, latency and packet loss have on effective WAN throughput. The measurements are used to approximate this metric on a per flow basis. While certain applications can take advantage of multiple flows for increased throughput, other applications (e.g. data replication) require high throughput on each individual flow.

WAN Bandwidth: 1024 Mbps

WAN Latency: 100 ms

WAN Loss: 0.5 %

Calculate

Maximum Effective Throughput (per flow): **1.65 Mbps**

Below are common latency and loss metrics:

LATENCY	
LAN	5 ms
Western US to Eastern US	80-100 ms
International	100-200 ms
Satellite	500+ ms
LOSS	
LAN	0%
Private Line	0%
MPLS	.1 to .5%
IP VPN	.5 to 1%

Why is this problem growing?

- Globalization
- Public Internet backbone
- RDBMS -> NoSQL, IMDG
- Rich Data Types
- Mobile Apps
- Client Side Data
- DR and BCP

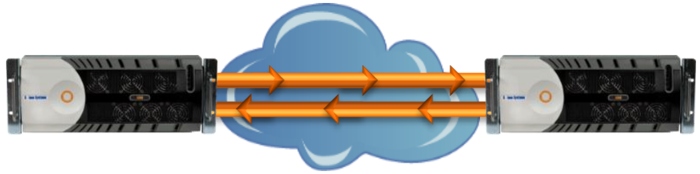
1. Bandwidth
2. Latency (RTT)
3. Error Rate (Loss)

Solace Systems™

Traditional WAN optimization techniques

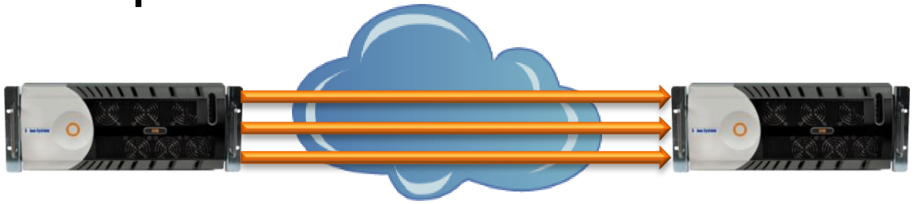
- Deduplication
- Compression
- Latency optimization
- Caching/proxy
- Forward error correction
- Protocol spoofing
- Traffic shaping
- Equalizing / Prioritizing
- Connection limiting
- Rate limiting

Bi-directional Message Streaming



Hardware Compression

Multiple Parallel Connections



Solace Systems™

Offloading the IP Stack to Hardware

Cavium Octeon II

- **32 core MIPS64 Processor**

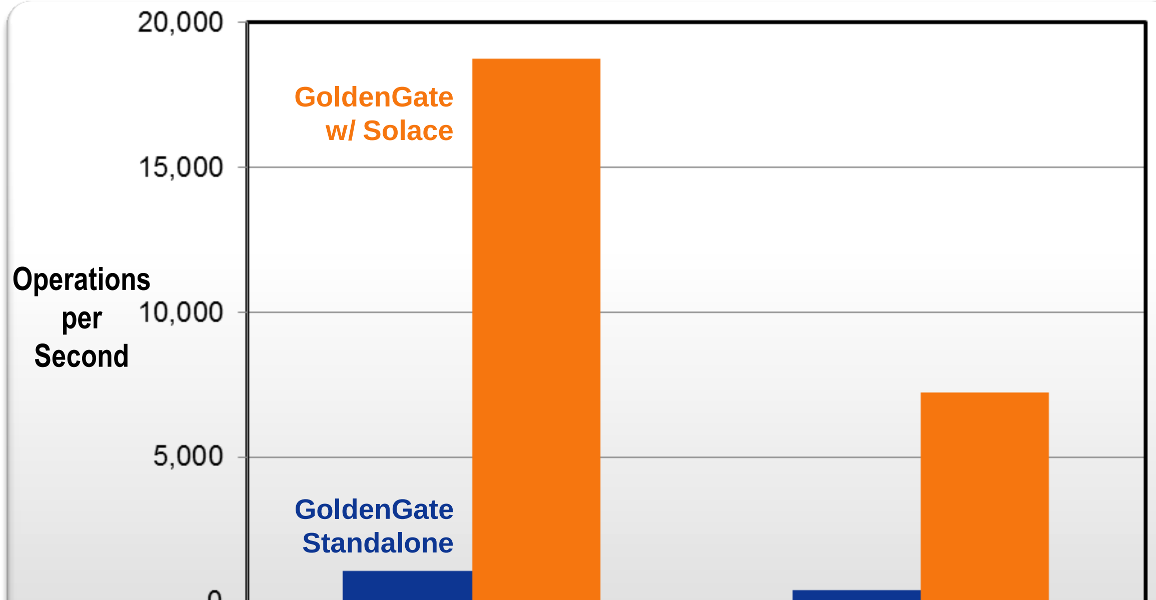
Pre-built application acceleration engines

- **Packet Processing**
- **Encryption/Decryption**
- **Deep Packet Inspection (RegEx)**
- **Compression/decompression**
- **De-duplication**
- **RAID**

Millions of concurrent connections

Improving the Speed of GoldenGate Synchronization

- **Real customer results**
  - Tested synchronization over a 622 Mbps link between New York and London with 75 ms round trip time
  - And over a 45 Mbps link between New York and Tokyo with 175 round trip time
- **Solace 18x Faster**



	New York to London 622 Mbps link, 75 ms RTT			New York to Tokyo 45 Mbps link, 176 ms RTT	
Operations Per Second	1,092	18,761		401	7,238
Transactions Per Second	82	1,409		30	543

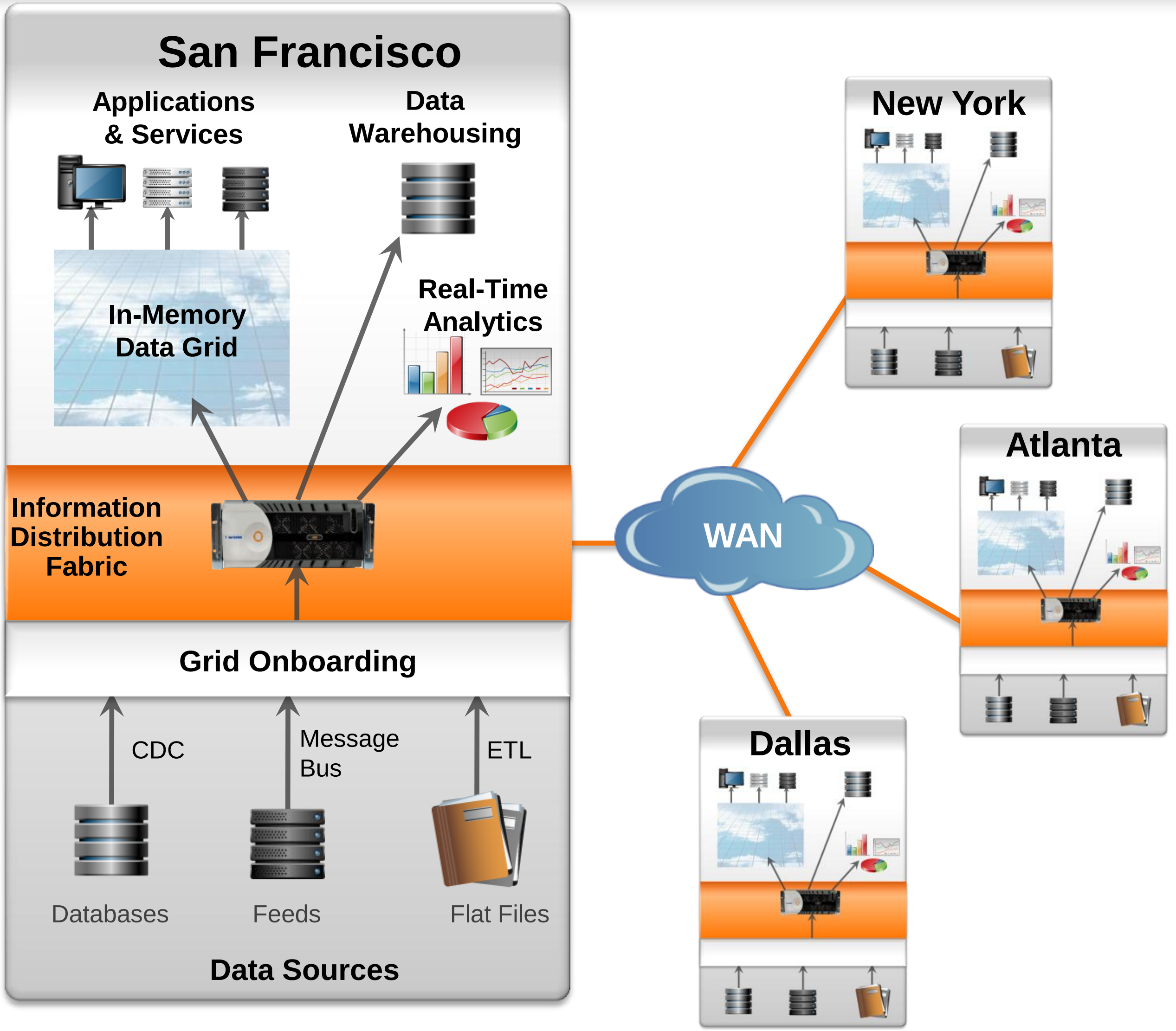
Back to the use case

Transactions != Packets

Database Records != Packets

Objects != Packets

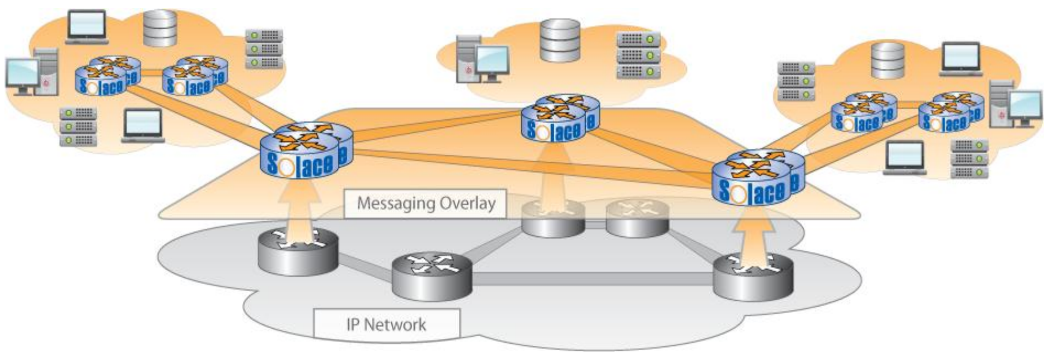
The Modern Information Distribution Fabric



Solace Systems™

Messaging Middleware Value

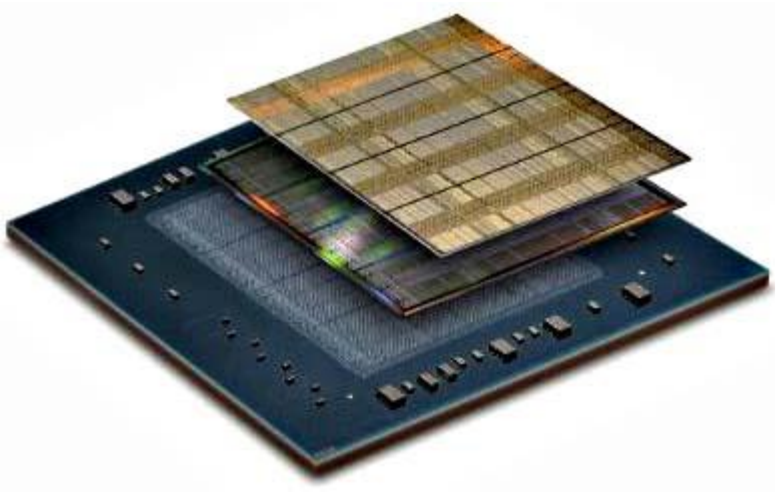
- Producer/Consumer Decoupling
- Disconnected Operation
- Location Independence
- Multipoint Delivery
- Advanced Filtering & Routing
- Message-based Granularity



Messaging layer on top of your IP network, so you can make messaging a shared optimized service.

Solace Systems™

FPGA



Xilinx Virtex-7 2000T FPGA

- **More than twice the capacity and bandwidth offered by the largest monolithic devices**
- **2 million logic cells (equivalent to 20 million ASIC gates)**
- **6.8 billion transistors**

Horizontal scalability on a single chip

Intel Westmere-EX

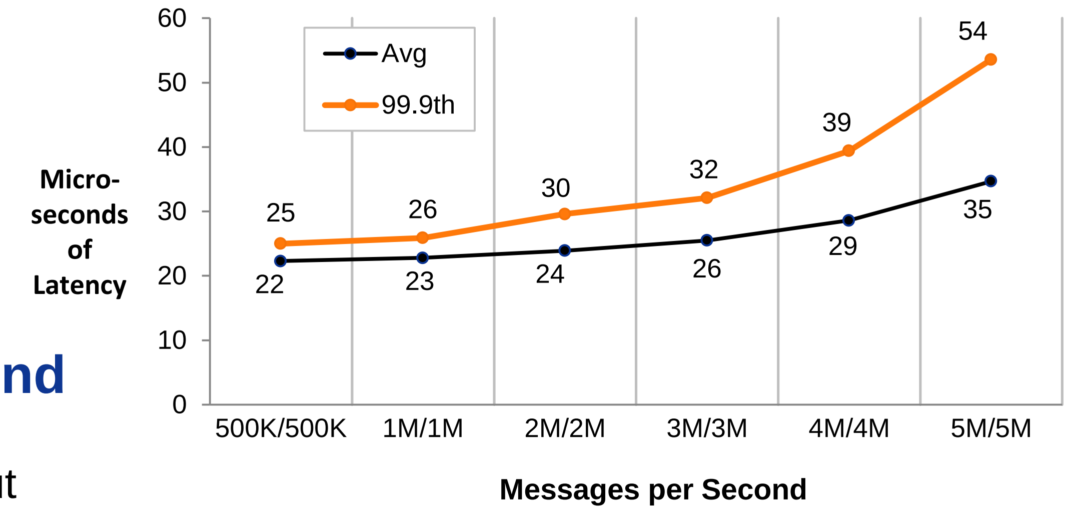
- **2.6 billion transistors**
- **10 64-bit cores @ 2.4 GHz**
- **7,200 MIPS**
- **130 watts TDP**

Xilinx Virtex-7 2000T FPGA

- **6.8 billion transistors**
- **3,600 8-bit processors @ 100 MHz**
- **180,000 MIPS**
- **20 watts TDP**

Reliable Messaging

- **Pure hardware solution**
  - No operating system
  - No context switching
  - No interrupts
  - No data copies
- **10 million messages/second**
  - Can be any combination, e.g. 5M in & 5M out, 2M in & 8M out



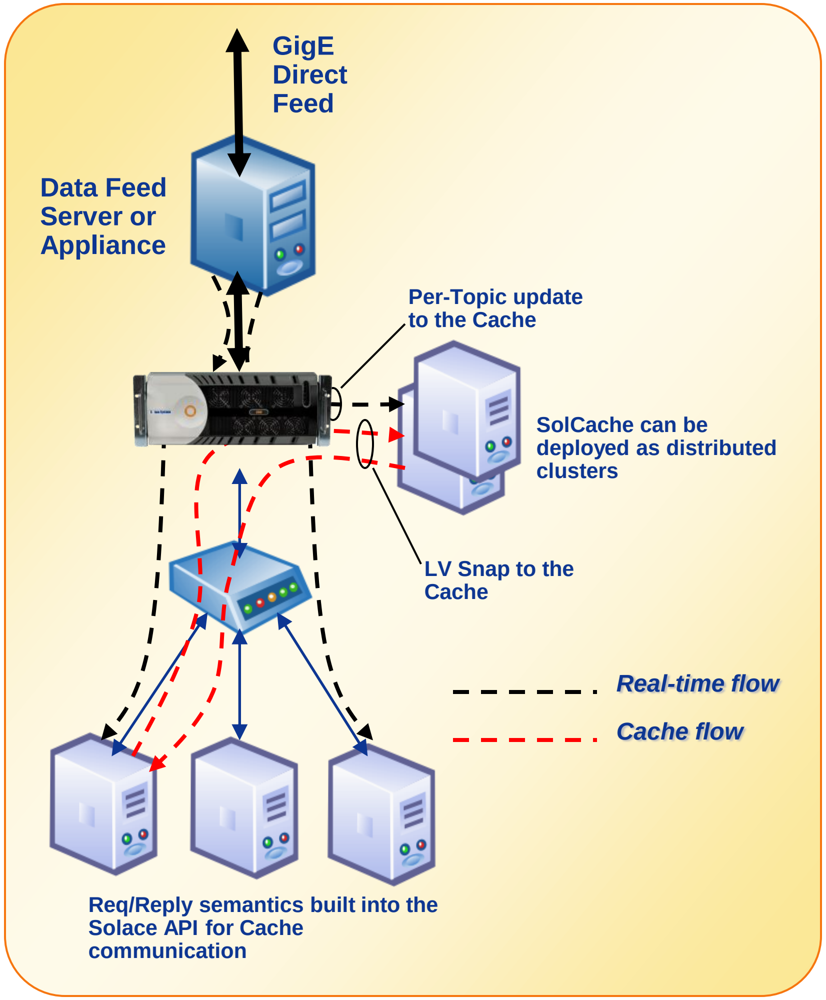
Bulk Message Rate	Message Size (bytes)	Message Rate (msgs/sec)	User Payload Bandwidth (Mbps)	
	100	5,930,000	4,744	
	500	2,080,000	8,320	*10GigE Line Rate the is Limit*
	1,000	1,080,000	8,640	
	12,000	92,000	8,832	
	30,000	34,000	8,160	

Solace Systems™

In Memory Message Caching

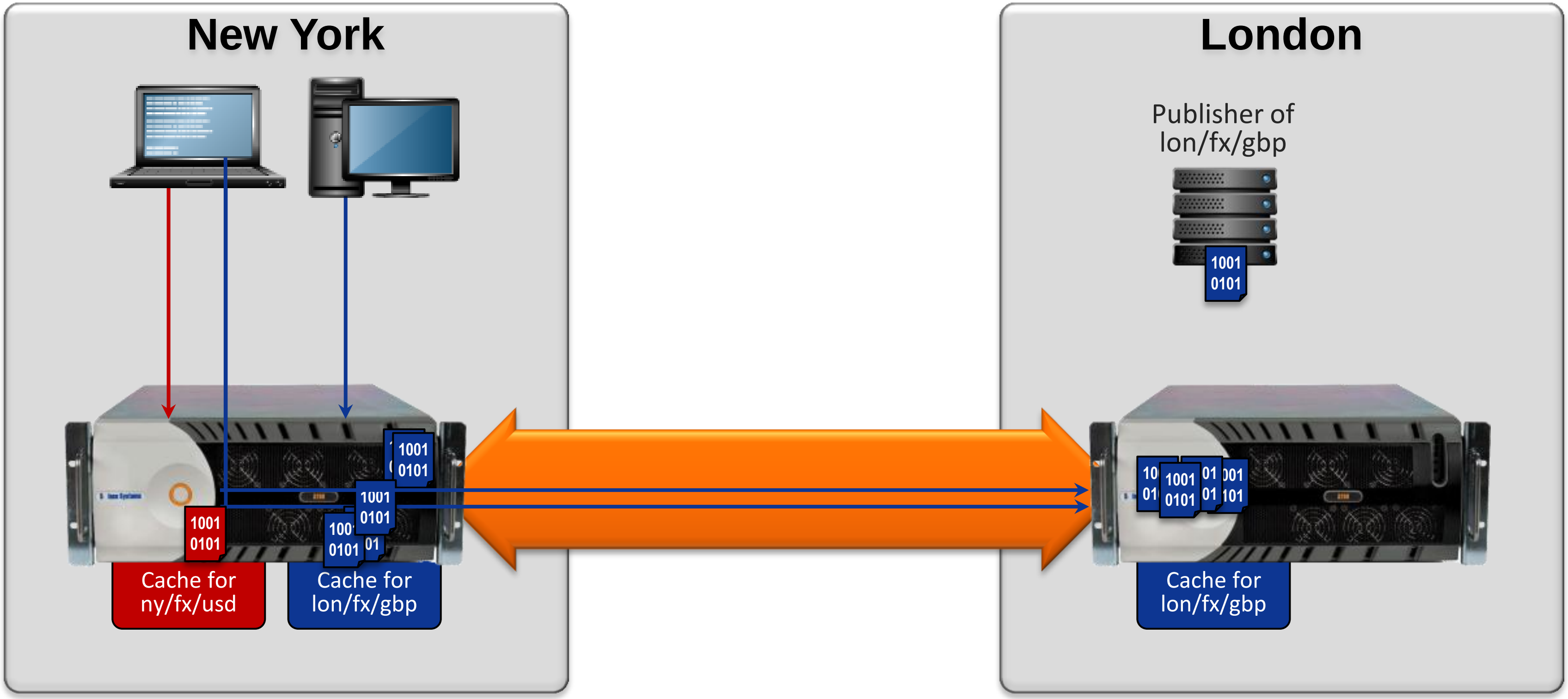
Non-persistent last-value cache can handle any payload

- Cache by number or timeframe
- Can run on appliance,
  or as a 64-bit app on a Linux server
- Centralized management of all caches.
- Clustering for load balancing and redundancy
- Support wildcard requests
- Request/reply with the cache, and control of synchronization of cache requests
- Topic names partitioned amongst instances to scale storage



Solace Systems™

Cascading Cache

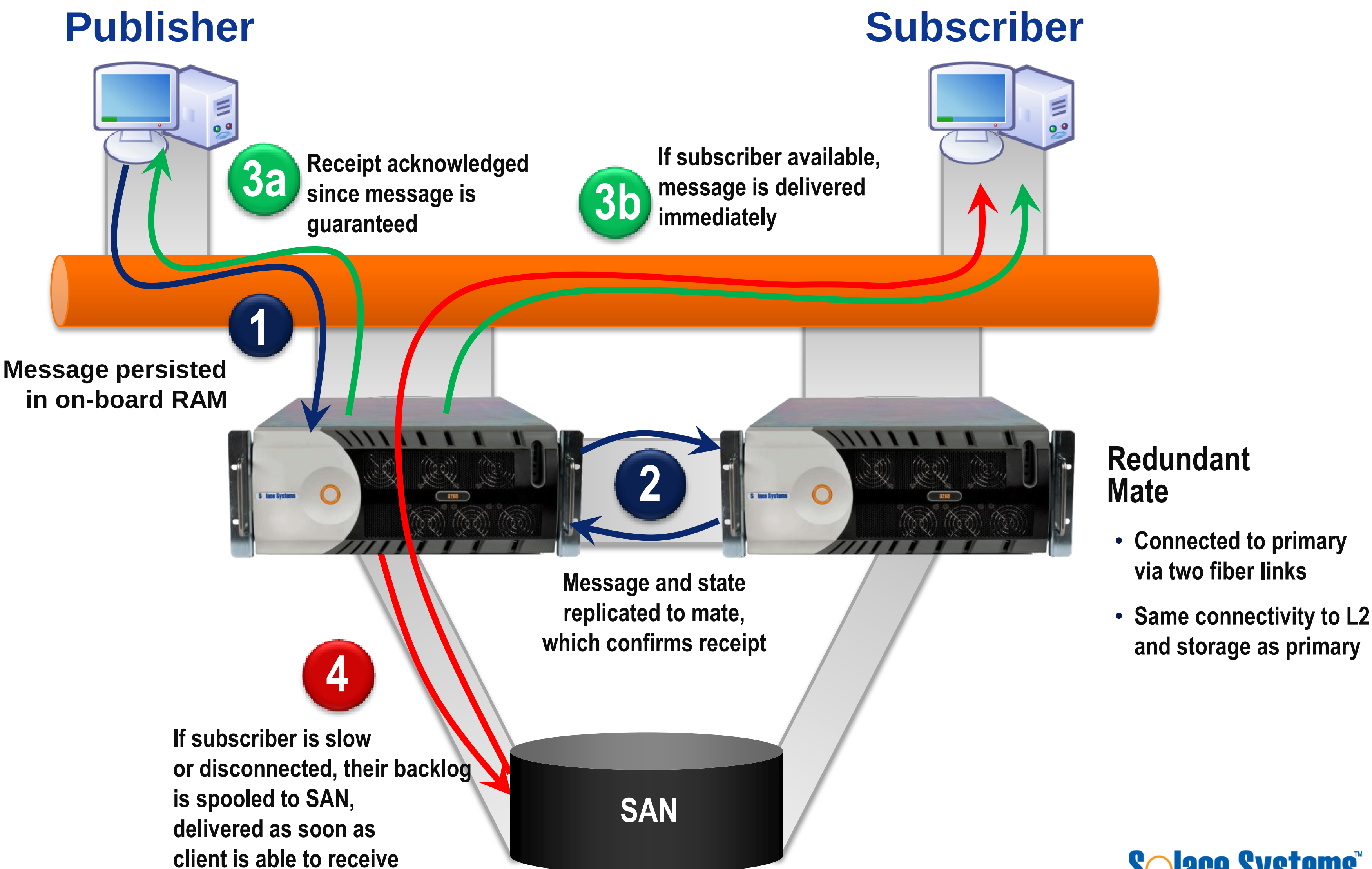


Incremental Updates

Cache Contents for NY/EQ/JNPR	Updated Cache Contents
SYMBOL: JNPR	SYMBOL: JNPR
VENUE: NYSE	VENUE: NYSE
LAST: 19.19	**LAST: 19.19**
VOLUME: 31,870	**VOLUME: 31,870**
DAY LOW: 19.03	DAY LOW: 19.03
DAY HIGH: 19.21	DAY HIGH: 19.21
52-WEEK LOW: 15.13	52-WEEK LOW: 15.13
52-WEEK HIGH: 23.98	52-WEEK HIGH: 23.98

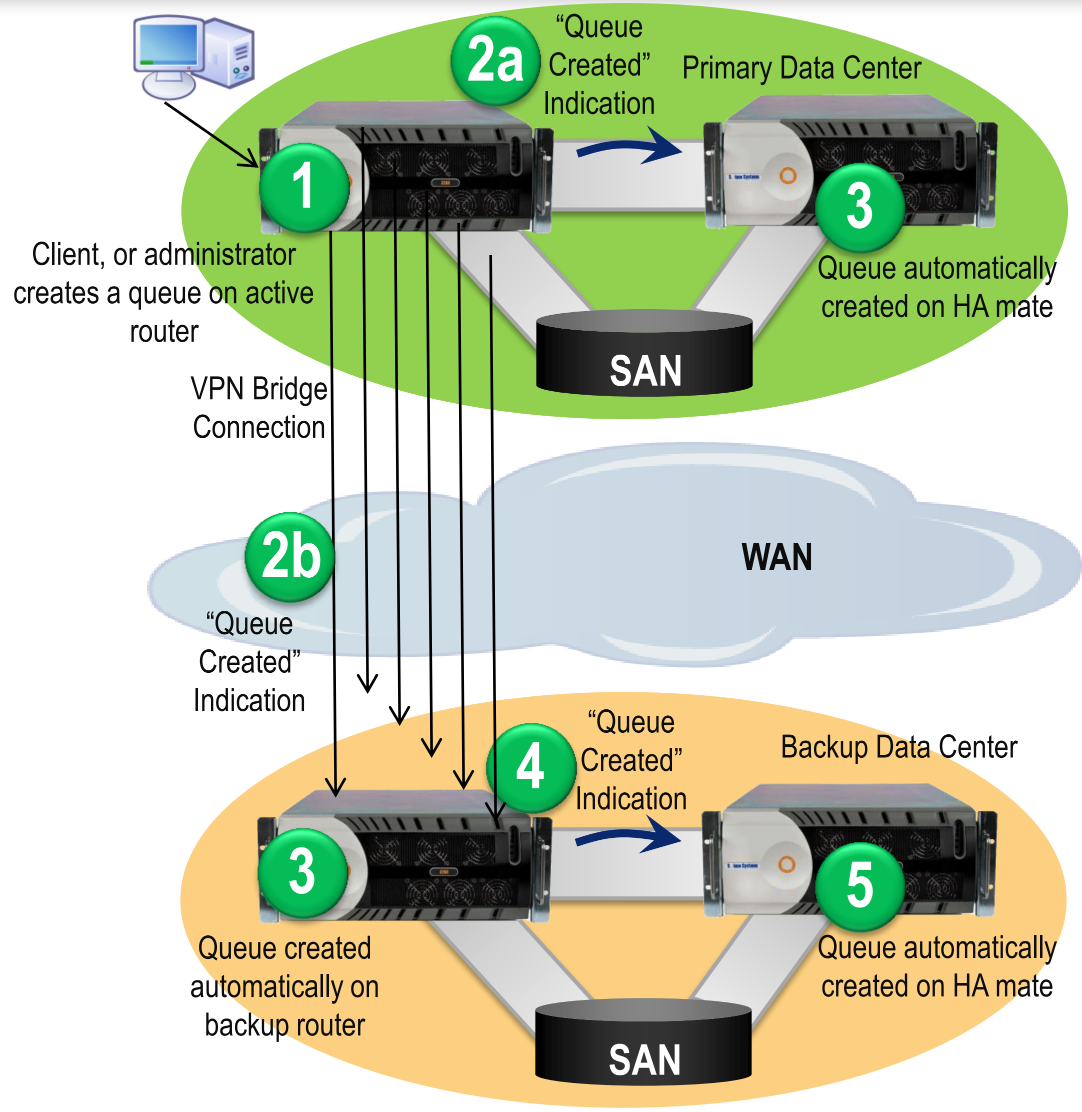
Hybrid Storage

Fault Tolerant Clustering of Messaging Nodes



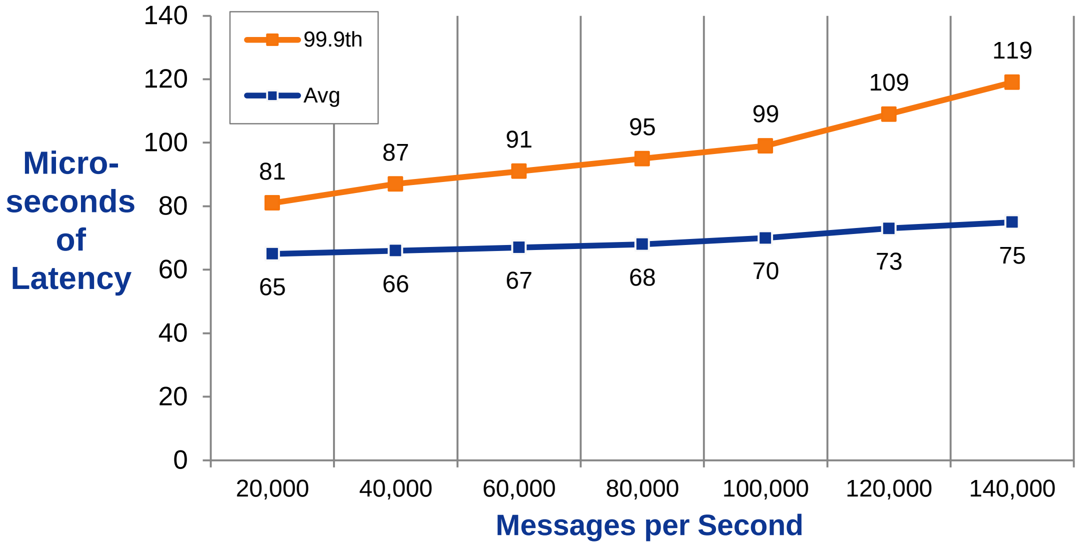
Multiple Datacenters and Disaster Recovery

- **Automatic replication of:**
  - Client-created endpoints
  - Configuration data
  - Transactional state
- **Needs Configurability for Sync & Asynchronous replication**



Guaranteed Messaging; Store & Forward Performance

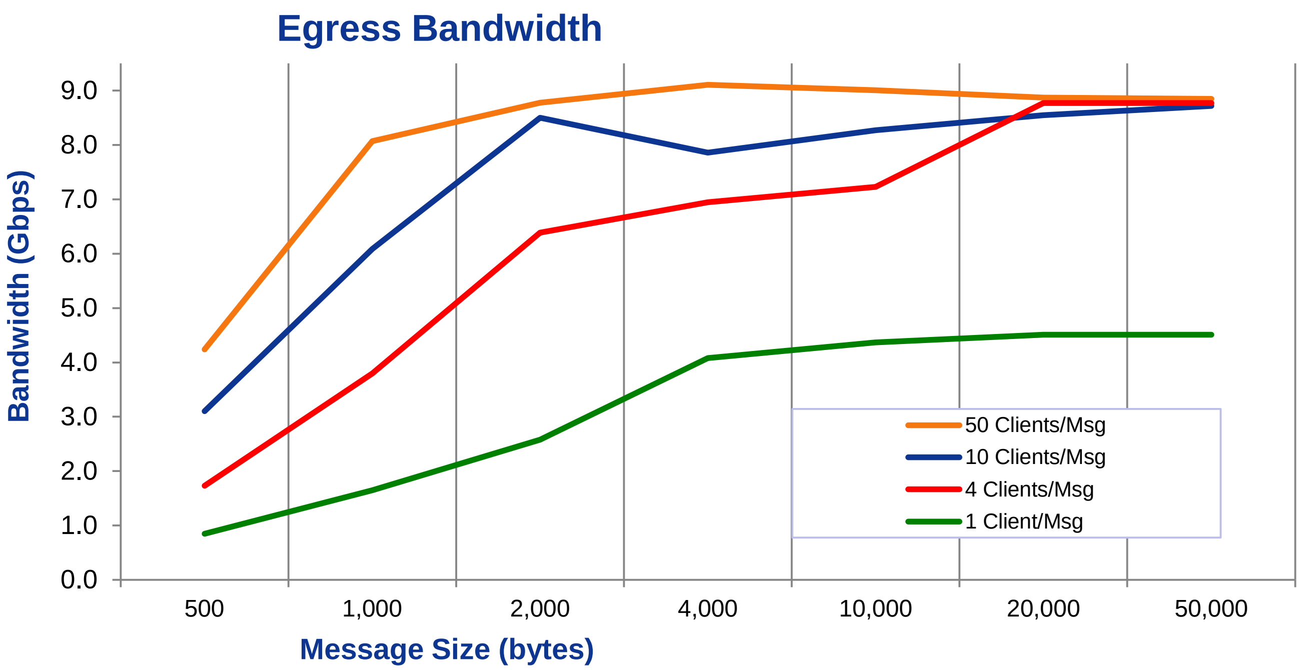
- **Failsafe w/o overhead of persisting every message to disk**
- **205K msgs/sec ingress and 205K msgs/sec egress**
- **Up to 4.5 Gbps of guaranteed messaging bandwidth**
- **Consistent latency even when servicing slow or recovering subscribers**



	Message Size (bytes)	Message Rate (msgs/sec)	User Payload Bandwidth (Mbps)
Bulk Message Rate	100	206,400	165
	512	206,400	845
	1,024	202,000	1,655
	2,048	157,500	2,580
	4,096	124,400	4,076
	10,240	53,400	4,375
	20,480	27,500	4,506
	51,200	11,000	4,506

Solace Systems™

Guaranteed Messaging; Fan-out performance

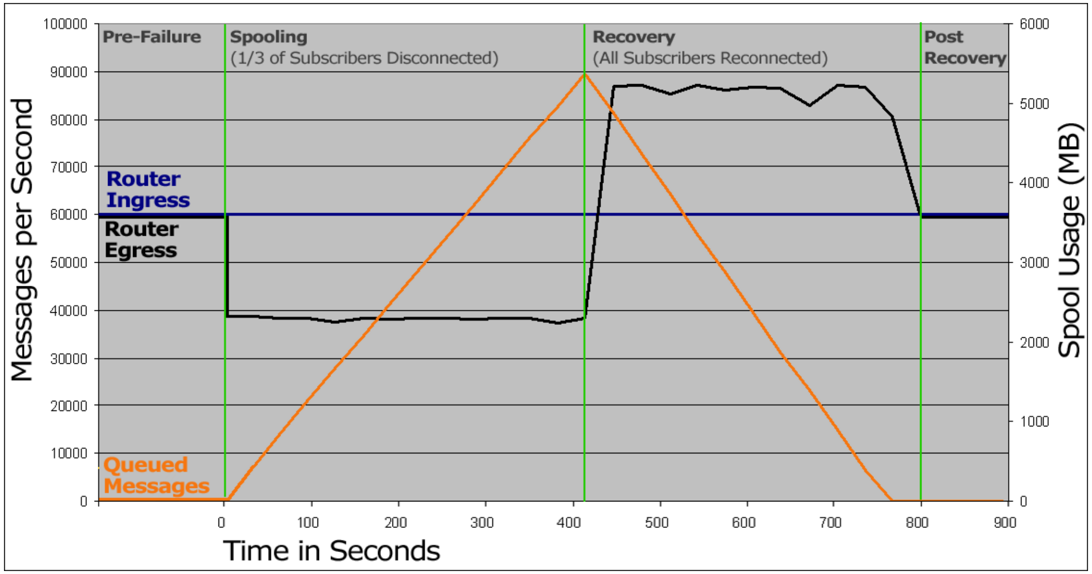


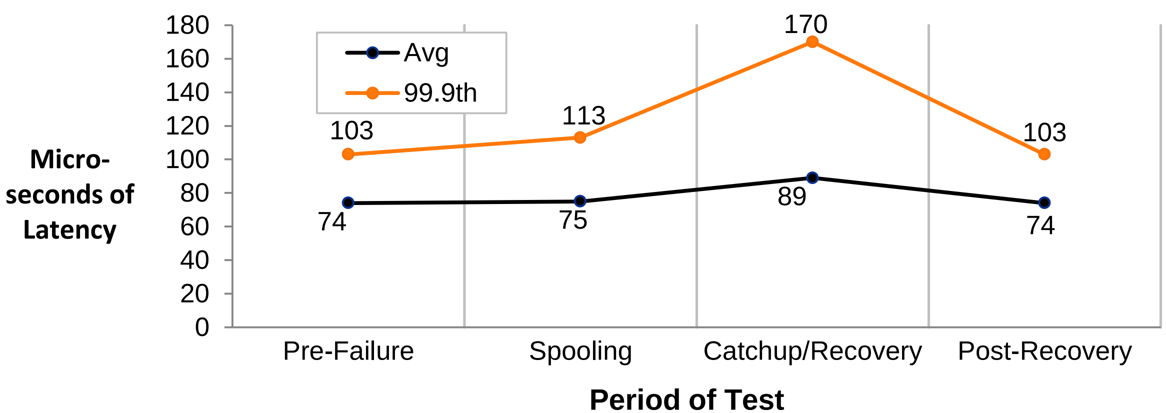
Message Size (bytes)	Egress Rate, 1 client/msg (msgs/sec)	Egress Rate, 4 clients/msg (msgs/sec)	Egress Rate, 10 clients/msg (msgs/sec)	Egress Rate, 50 clients/msg (msgs/sec)
512	206,400	422,000	756,000	1,035,000
1,024	202,000	464,000	744,000	985,000
2,048	157,500	390,000	519,000	536,000
4,096	124,400	212,000	250,000	278,000
10,240	53,400	88,300	101,000	110,000
20,480	27,500	53,500	52,200	54,150
51,200	11,000	21,400	21,300	21,600

Solace Systems™

Offline or Slow Consumer Handling

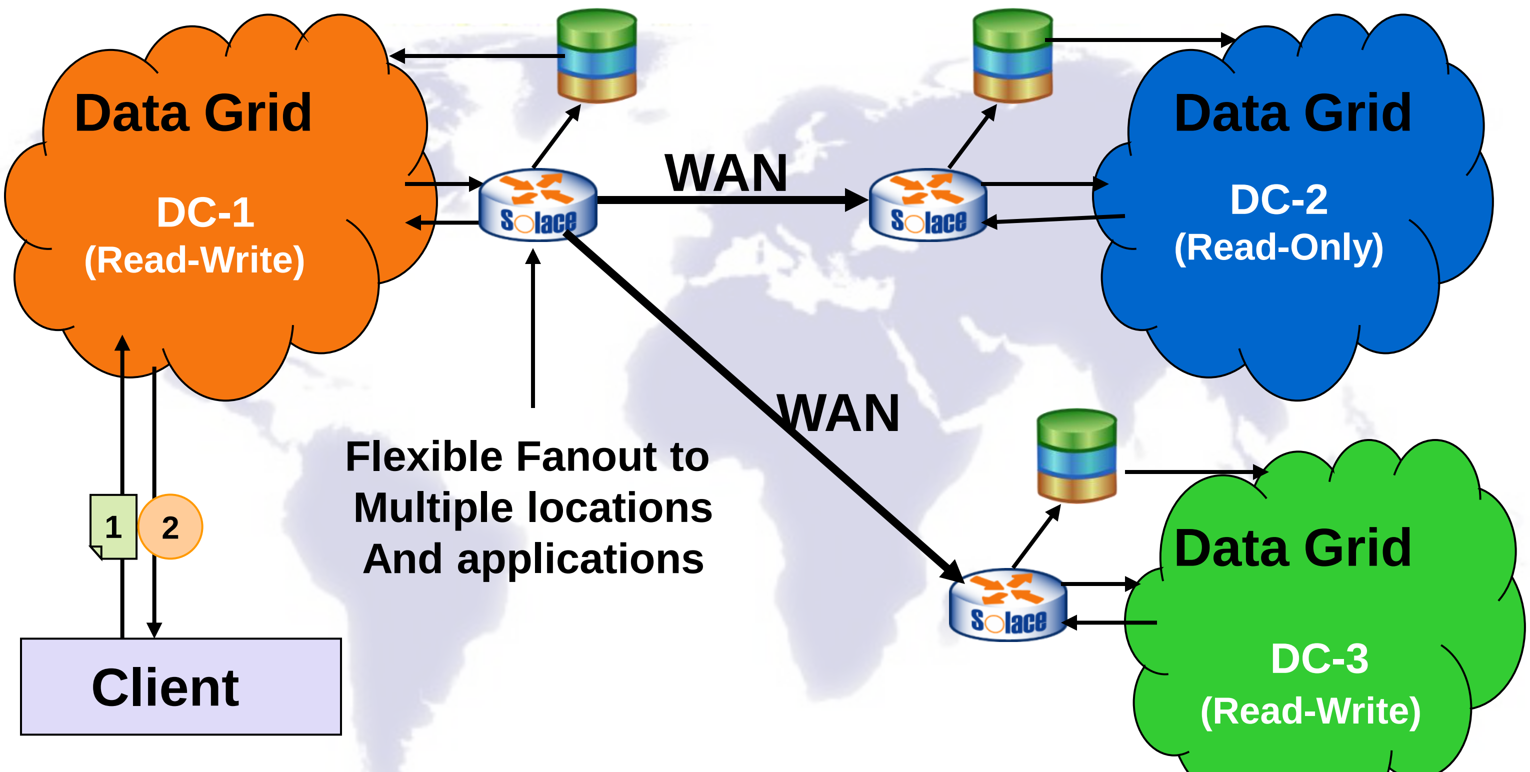
- Publisher rates not affected by slow/offline consumers
- Fast consumers not affected in rate or latency by slow/offline consumers
- Re-connected subscribers “catch up” without impacting other clients
- Behavior & performance cannot be matched by software due to patented technology





Solace Systems™

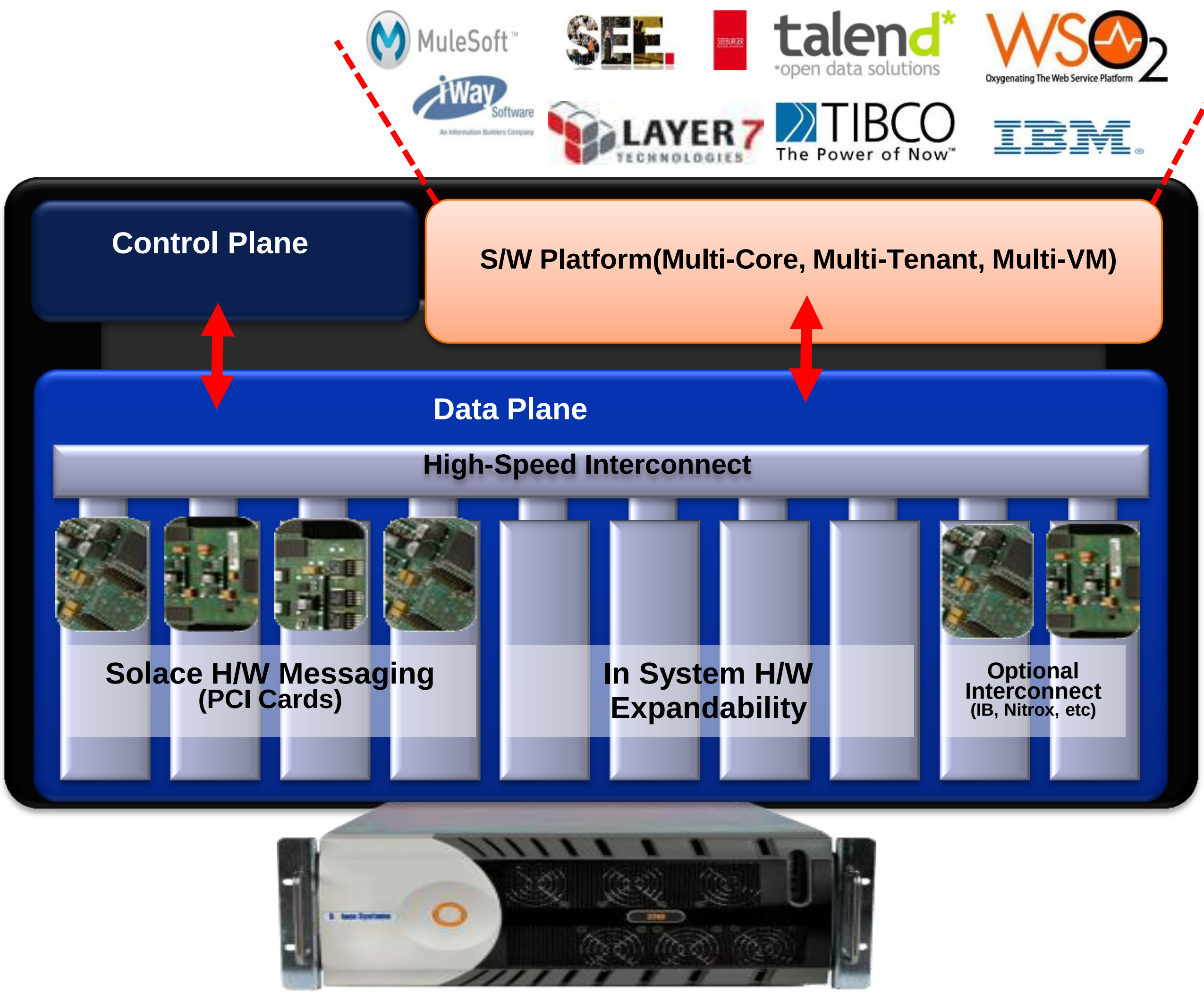
Optimized In-memory Grid Replication



Single Infrastructure for Grid Synchronization:

- inherently one-to-many, so can propagate
  to many other sites/instances – either locally or over the WAN
- Supports DR, Active/Passive, or Active/Active architectures

Solace as an Appliance Platform



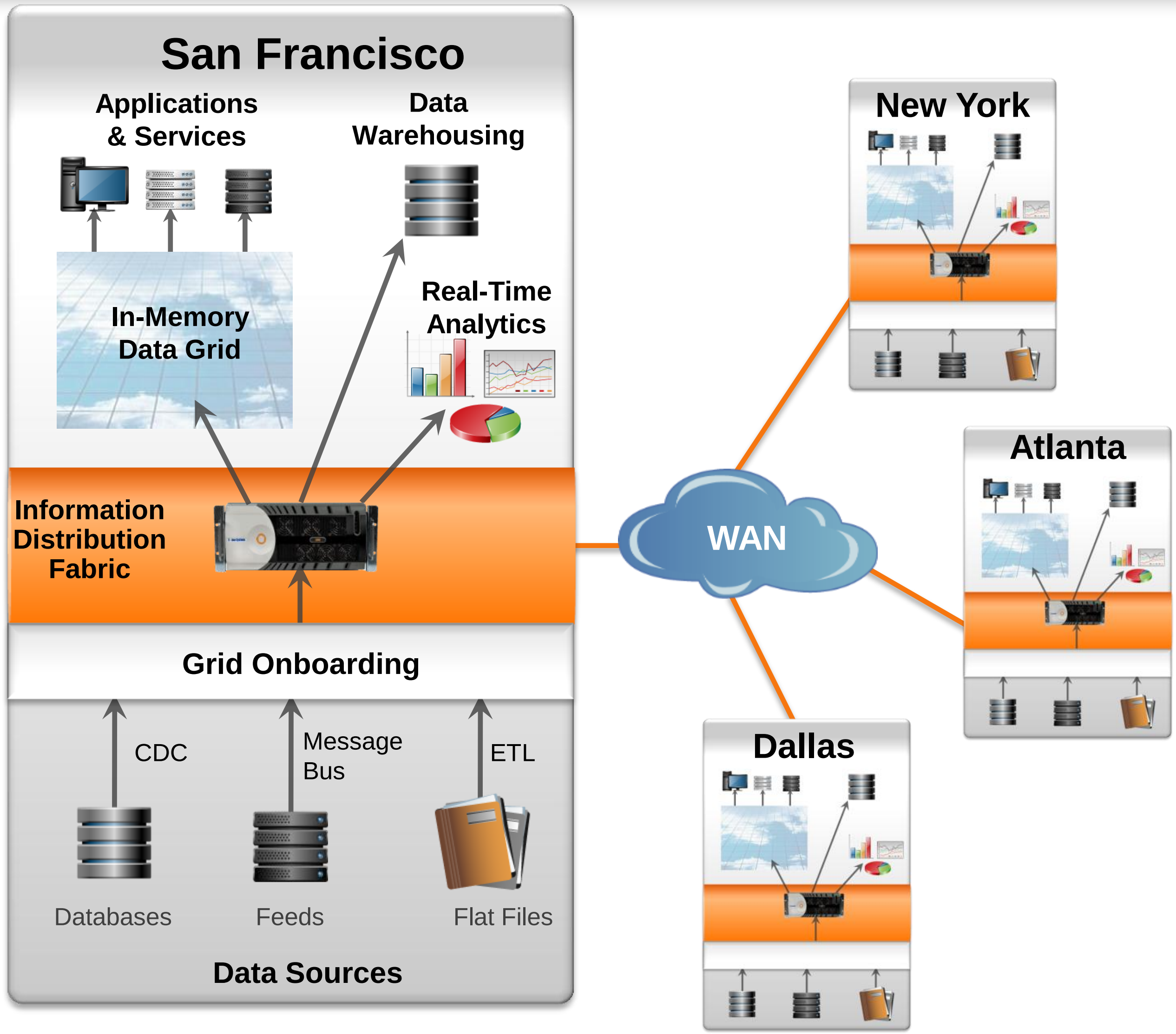
- Messaging all in hardware
- General purpose processors used to run 3rd party software that interacts seamlessly with hardware messaging internally
- Integration is easy with JMS
- Enables flexible solution options within an appliance

APIs: JMS,C, .Net, Java, JavaScript, Flash, Silverlight, iOS, Node.js, Ruby, Python, etc.

The Modern Information Distribution Fabric

High Volume Onboarding

Fast, Efficient WAN Sync



Solace Systems™