12th ANNUAL WORKSHOP 2016

INFINIBAND AS CORE NETWORK IN AN EXCHANGE APPLICATION

R. Barth, Applications & Architecture
J. Stenzel, Systems Infrastructure

Deutsche Börse AG

April 5th, 2016

OVERVIEW

Eurex Exchange's T7 is based on Deutsche Börse Group's proprietary global trading architecture, which is also in use at the International Securities Exchange (ISE) and the Bombay Stock Exchange.

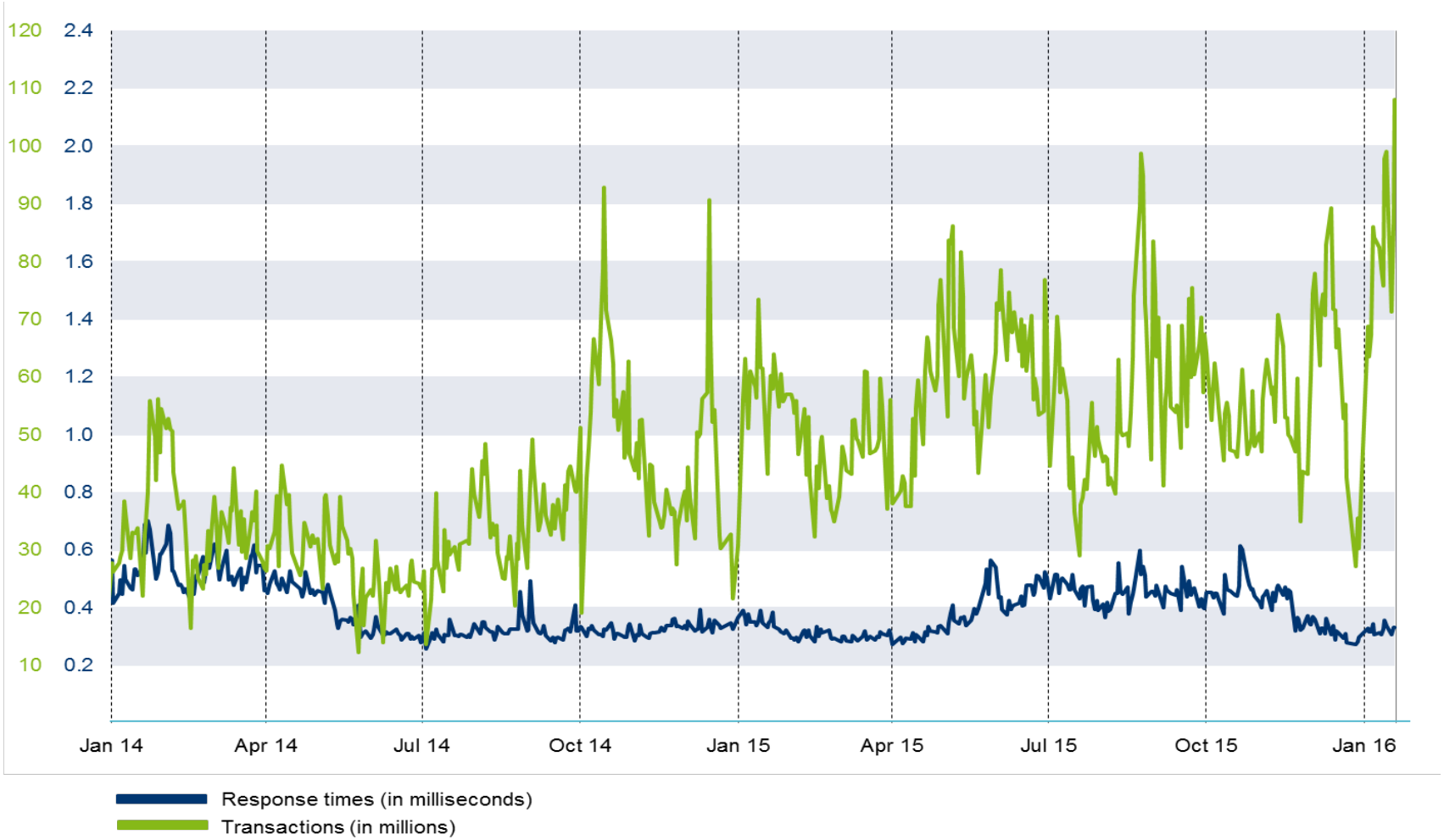
This presentation will present the network relevant part of the application design and its interdependence with the combination of InfiniBand and IBM WebSphere MQ Low Latency Messaging (LLM).

For further details about T7 please have a look at our web page:
www.eurexchange.com/t7
www.ise.com/technology/t7/

DESIGN REQUIREMENTS

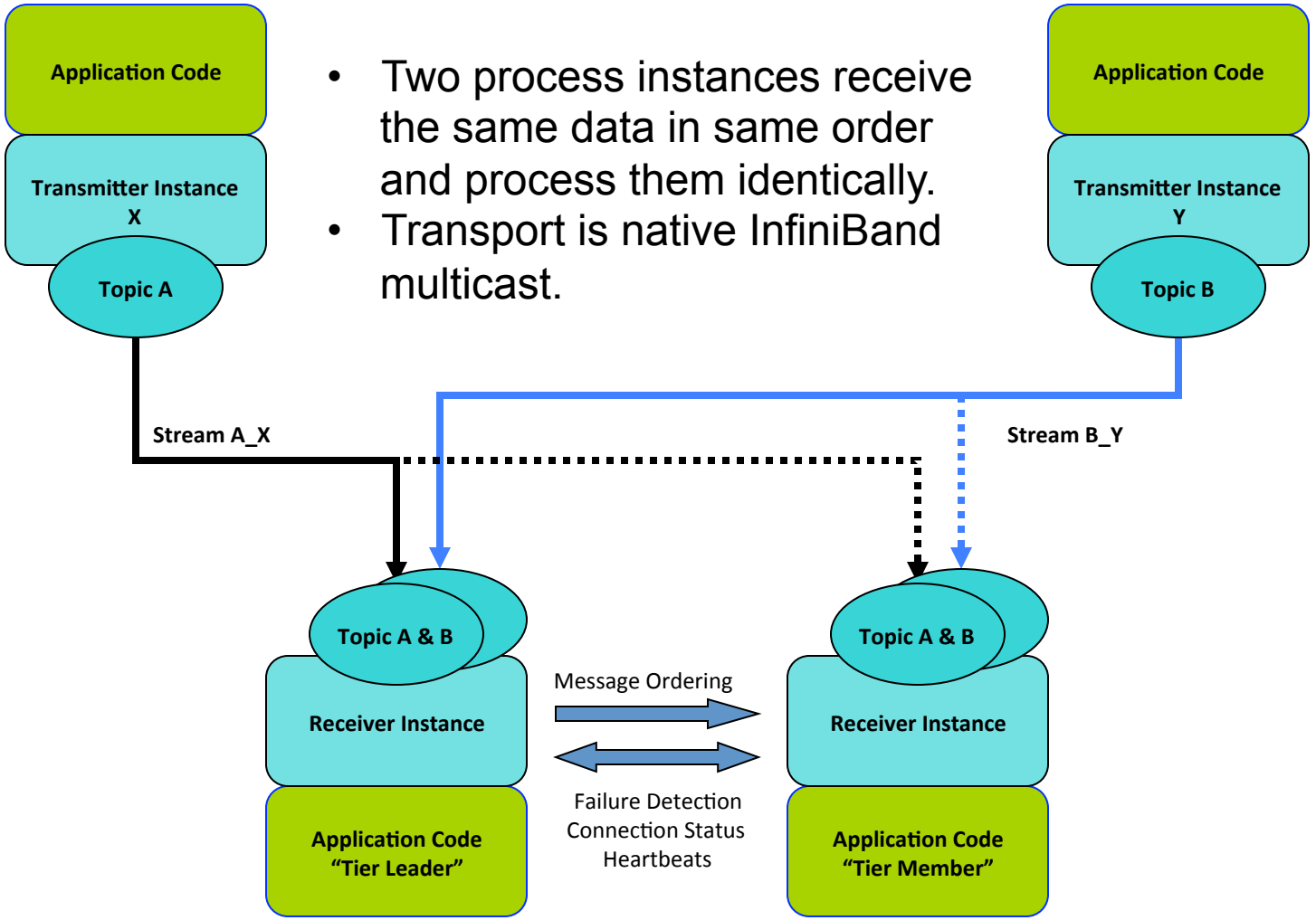
- **Low latency**
  - Fast response to the members independent from the system load
- **High throughput**
  - Number of transactions which can be processed in peak and in sustained rate without queuing
- **Consistent latency and throughput**
  - Not only the average numbers are relevant, also the minimum, median and maximum (99%) are significant
- **High availability**
  - A trading system provides an infrastructure for the financial markets
  - Has to fulfill regulatory requirements in sense of availability

TRANSACTIONS & RESPONSE TIMES

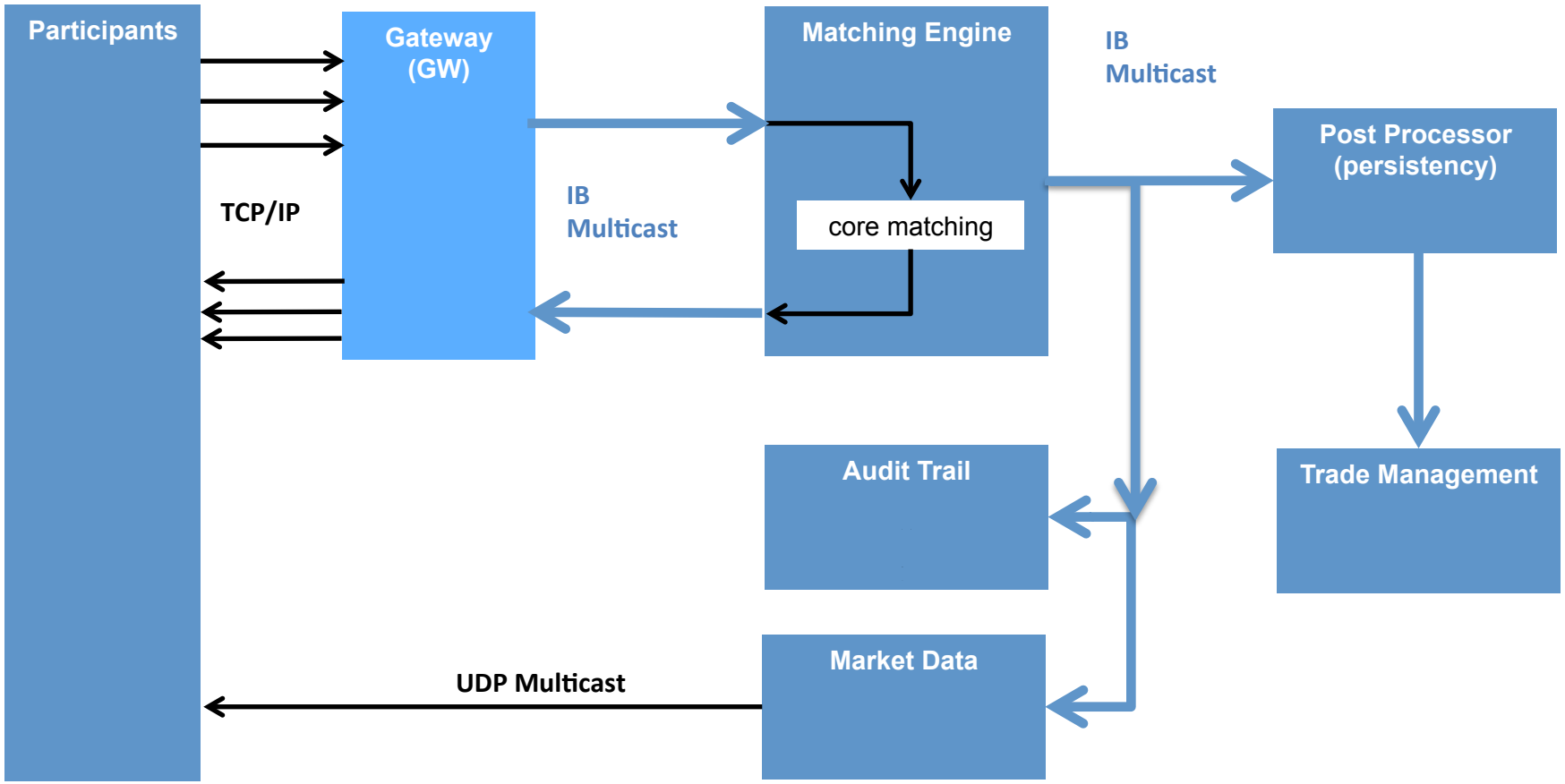


LLM HOT-HOT TIER COMMUNICATION

- Two process instances receive the same data in same order and process them identically.
- Transport is native InfiniBand multicast.



EUREX T7 TOPOLOGY

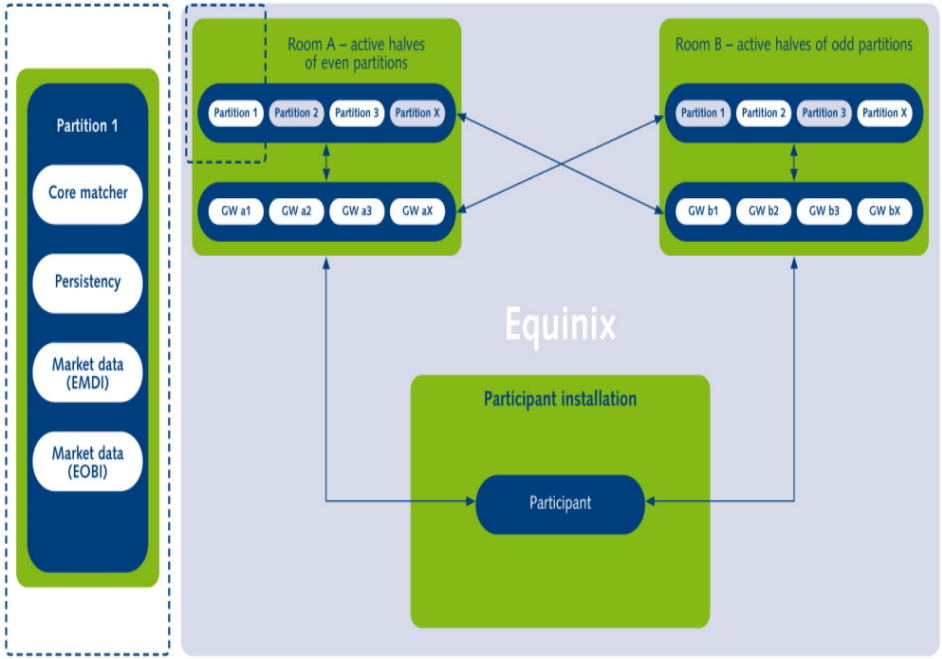


EUREX T7 PRODUCTION OVERVIEW

- T7 consists of partitions. Here, a partition is a failure domain in charge of matching, persisting and producing market data for a subset of products. Each T7 partition is distributed over two rooms in the Equinix data centre.
- There are 10 T7 partitions in charge of futures and options trading available on Eurex Exchange. A separate additional partition is used for European Energy Exchange (EEX) products.
- There are 16 high-frequency gateways in the Equinix data centre shared by all Trading Participants of Eurex Exchange.
- The reference data contains the mapping of products to partition IDs. The physical location of the Eurex Enhanced Trading Interface gateways in the Equinix data centre relative to the room where a matching engine resides has no impact on the order latency.

- Note that normally the active half of a partition is either in room A (for even partitions) or in room B (for odd partitions).
- Only in case of the failure of a matching engine or a market data publisher, the active half of the service will shift to the other room.

T7 data center topology



WHY MULTICAST

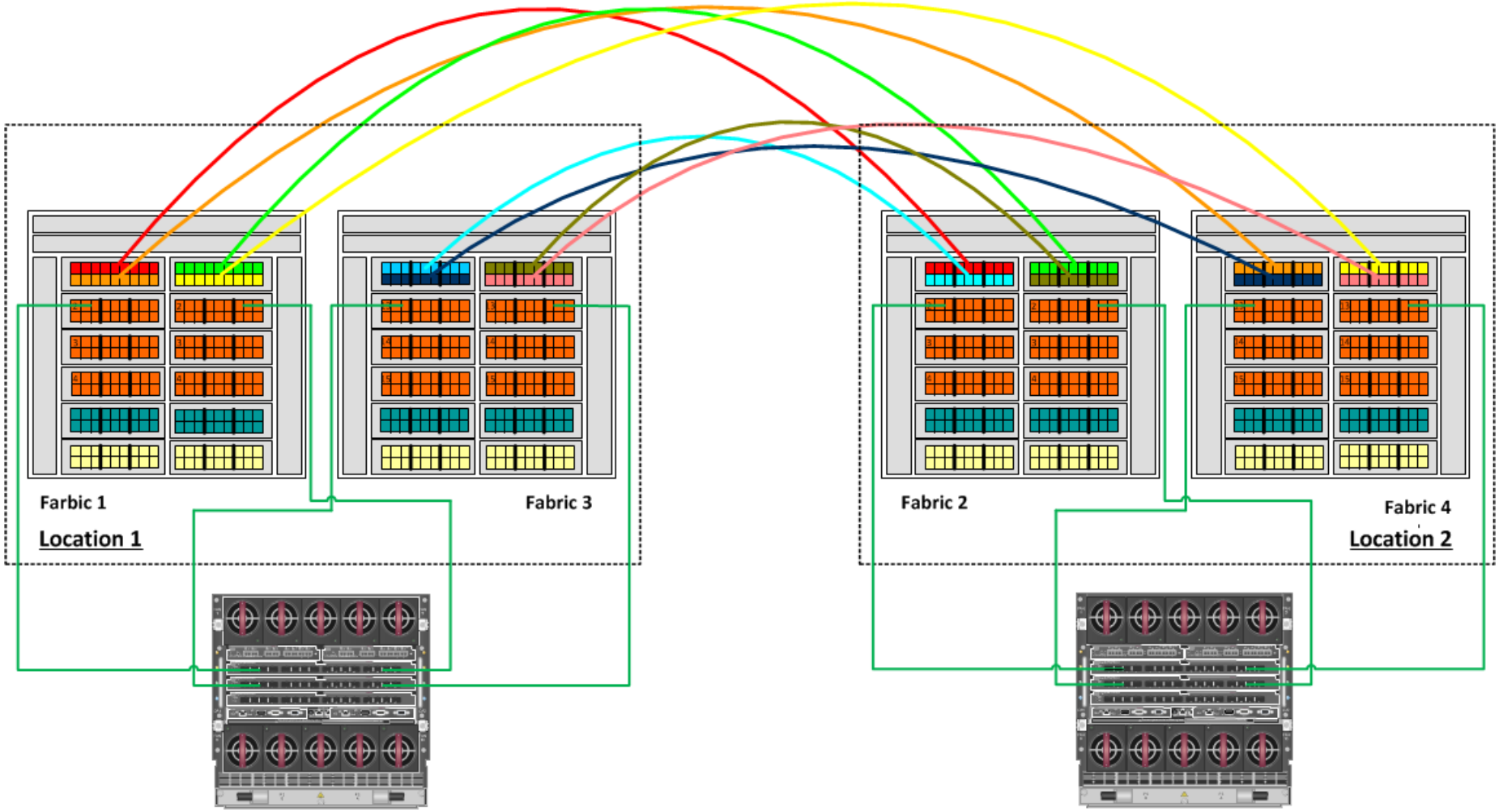
- **LLM hot-hot tier concept requires delivery to several receivers (tier members)**
- **Simultaneous delivery to all receivers provides better fairness as there is no order and delay due to single send calls**
- **Enforces functional layouts that avoid single addressing patterns in the core, reduces encoding/message generation overhead as only a single message is required**

MULTICAST GROUPS & NODES

- **ISE (3 Markets)**
  - 410 nodes
  - approx. 750 native IB multicast groups
- **EUREX**
  - 200 nodes
  - approx. 310 Native IB multicast groups

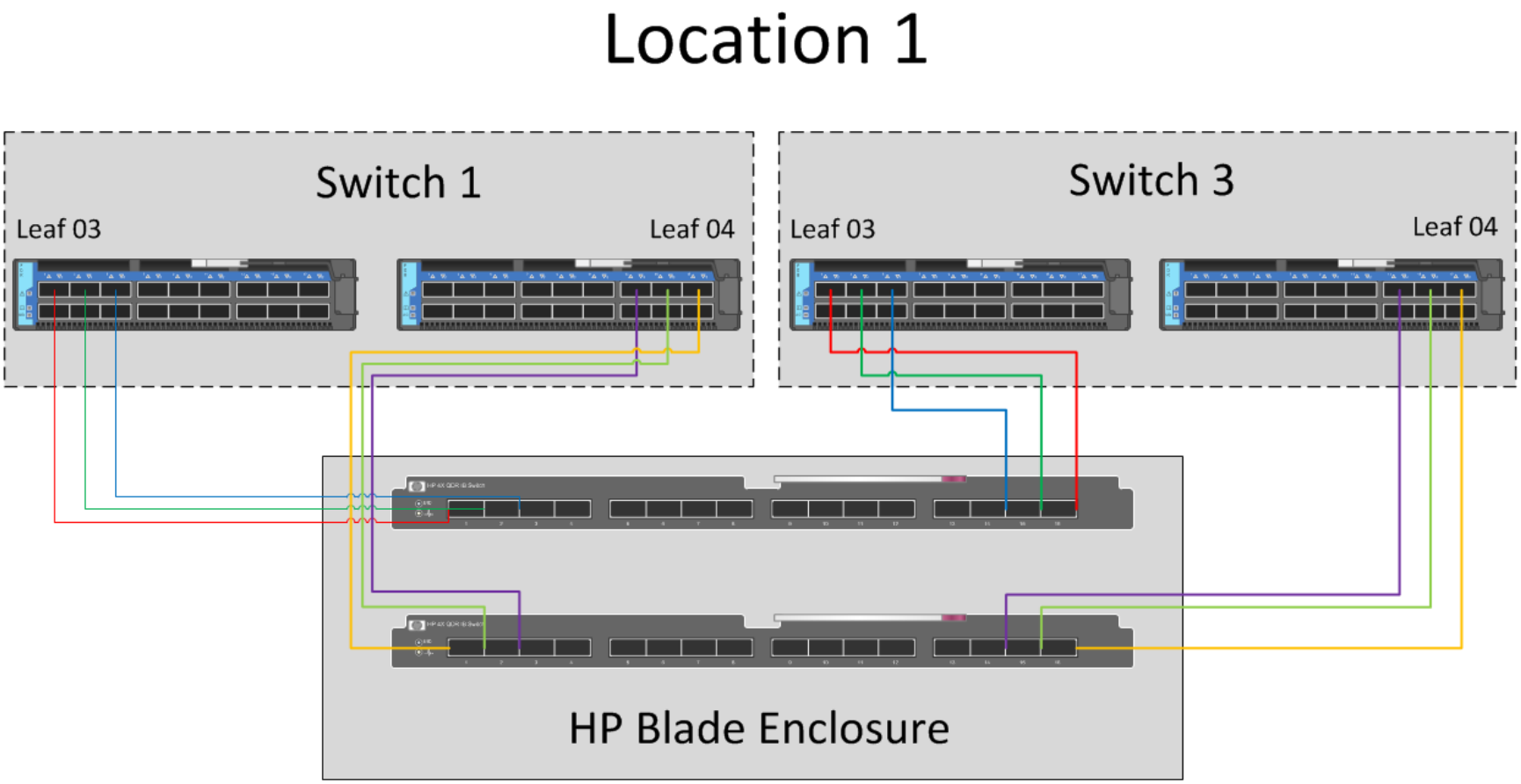
Number of groups/mlids scales with number of partitions times number of gateways.

CROSSOVER CABLING FOR ROOT SWITCH



ENCLOSURE CABLING CONCEPT

Enclosure switches to fabric switches



LEAF MODULE USAGE

- **L01 and L02** crossover connection to each fabric
- **L03 to L08** connection to the enclosures
- **L08 to L12** connection to the rackmount server

INSTALLED CABLES

- 45 Installed cabling thru the fire protection wall between location 1 and 2
  - 36 cables are in use
  - 9 cables are spare
- 12 Cables per enclosure (6 cables per HP switch)
- 2 Cables per rackmount server (1 card with 2 ports)

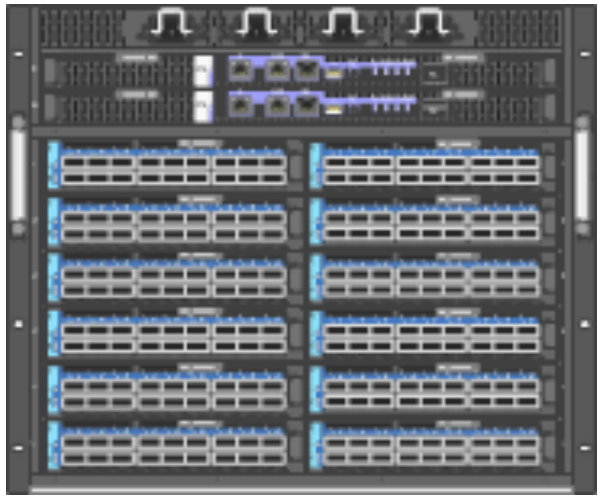
SERVER HARDWARE

- **HP BL460c Gen8**
  - MT27500 [ConnetX-3]
  - Flex LOM
  - 96G RAM
  - 900G hard drive (2 x 900G, 1 mirror set)
  - Fibre channel Qlogic (not all BL460 server)

- **HP DL380p Gen 8**
  - MT27500 [ConnetX-3]
  - Gigabit Ethernet
  - 96G RAM
  - 900G hard drive (4 x 900G, 2 mirror sets)
  - Fibre channel Qlogic (not all DL380 server)
  - Solarflare card for PTP (Precision Time Protocol)

4 x IS5200 InfiniBand Fabric

- 2 admin boards
- 12 leaf board
- 6 spine boards



IB ROUTING

Routing protocol: updn

Root switch: Fabric 1 L01 and L02

Fabric 4 L01 and L02

/etc/rdma/root_guid.conf

0x2c90200468808

0x2c90200478620

0x2c90200478658

0x2c902004786d8

LOW LATENCY CABLING CONCEPT

For the low latency concept we divided the server into 2 groups

Non critical servers

- No matching relevant server (Monitoring, Audit Trail, Trade Manager)
- 1 hop more for blades (enclosure switch)
- The non critical server are not close to the switches and connected with 10m cables

LOW LATENCY CABLING CONCEPT

Critical servers

- matching path relevant systems (High Frequent Gateway's, Matching Engine, Market Data)
- those Rackmounts servers connected with 5m cables
- location close to the Fabric, less latency
- fewer hops

MULTICAST ROUTING OBSERVATIONS

We observe sporadically outliers in latency. As we have no hardware time-stamps like in Ethernet and also not full transparency due to the usage of LLM we tried to investigate on network layer using InfiniBand port counters. Open questions and ongoing analysis are:

- **Already reduced the number of virtual lanes to 3 (analysis ongoing)**
- **Counters are quite often in overflow, so periodically resets must be implemented (not yet done)**
- **Multicast forwarding tables show, that only one port is used for switch to switch routing.**
  - Do we have congestion at that point?
  - Are there better cabling strategies for Multicast dominated networks
  - Better algorithms as manual GUID routing orders may not help for multicast
  - ……

OPENFABRICS
ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

Ralph Barth & Joachim Stenzel

Deutsche Börse AG