Shared Memory Communications over RDMA (SMC-R) Performance update: SMC-R over distance

Gus Kassimis – kassimis@us.ibm.com

September 2014

SMC-R and RoCE performance at distance

- Initial statement of support for SMC-R and RoCE Express
  - 300 meters maximum distance from RoCE Express port to 10GbE switch port using OM3 fiber cable
    - 600 meters maximum when sharing the same switch across 2 RoCE Express features
    - Distance can be extended across multiple cascaded switches
    - All initial performance benchmarks focused on short distances (i.e. same site)
- Updated testing for RoCE and SMC-R over long distances
  - IBM System z™ Qualified Wavelength Division Multiplexer (WDM) products for Multi-site Sysplex and GDPS ® solutions qualification testing updated to include RoCE and SMC-R. Two vendors already certified their DWDM solution for SMC-R and RoCE Express:

    1. Fibernet DUSAC 4800 Release 2.2b - on two client cards, the FTX-n and the FTX-10C (both cards are single port transponders). The qualification letter for this release can be found at the following link:

       https://www-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/FibernetSL?OpenDocument&pathID=

    2. Cisco 15454 Release 9.6.0.5 - on the 10 x 10G client card (15454-M-10x10G-LC) in 5:5 transponder mode.The qualification letter for this release can be found at the following link:

       https://www-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/ciscoSystemsInc?OpenDocument&pathID=

- *But how does SMC-R and RoCE perform at distance?*

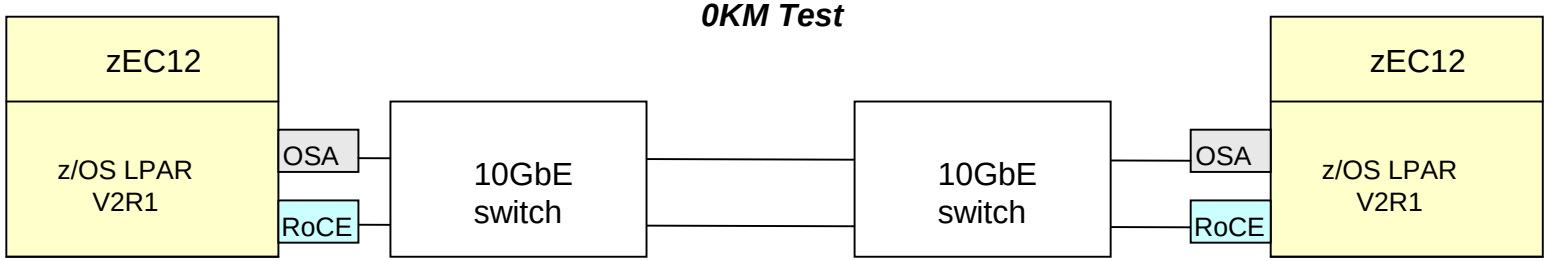
Internal IBM testing with SMC-R and RoCE at distance

- Performance benchmarks performed using the IBM Application Workload Modeler (AWM) tool
- Micro-benchmarks: Tests included AWM in client and server mode on separate z/OS LPARs generating TCP socket traffic
  - No business logic in AWM (simply sends/receives data)
  - Does exercise full TCP/IP API and protocol stack layers
- Environment: 2 z/OS LPARs on zEC12 with 2 dedicated CPUs each with following connectivity
  - OSA Express 5S (For TCP/IP benchmarks)
  - RoCE Express (For SMC-R benchmarks)

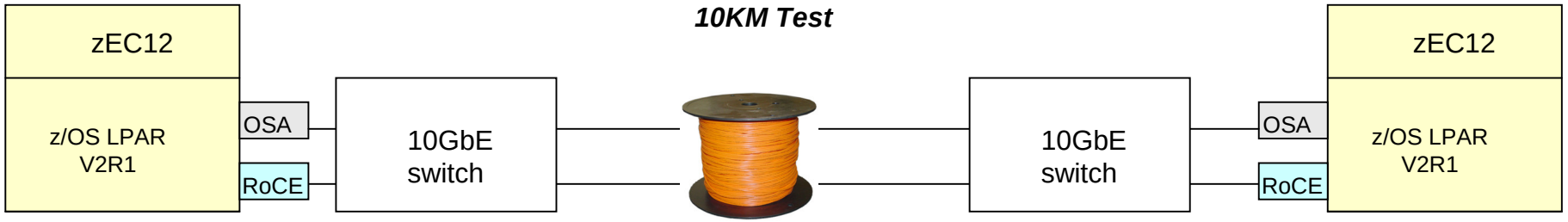
- ***NOTE:*** Based on internal IBM benchmarks using a modeled socket workload in a controlled laboratory environment using micro benchmarks. Your results may vary based on your configuration, workloads and environment.

Internal IBM testing with SMC-R and RoCE at distance - Configurations

0KM Test

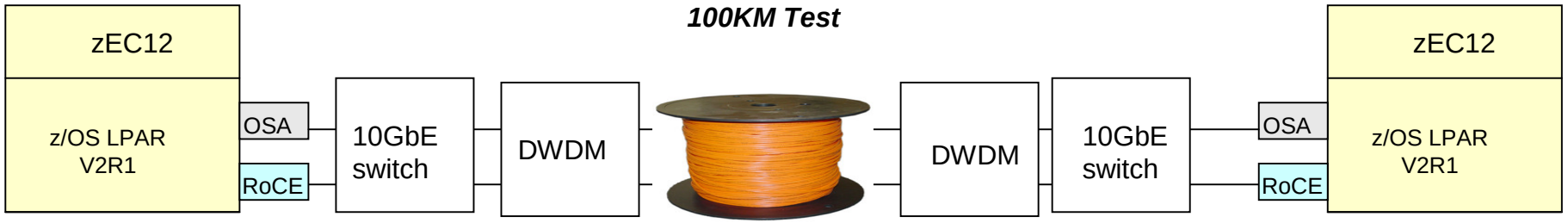


10KM Test



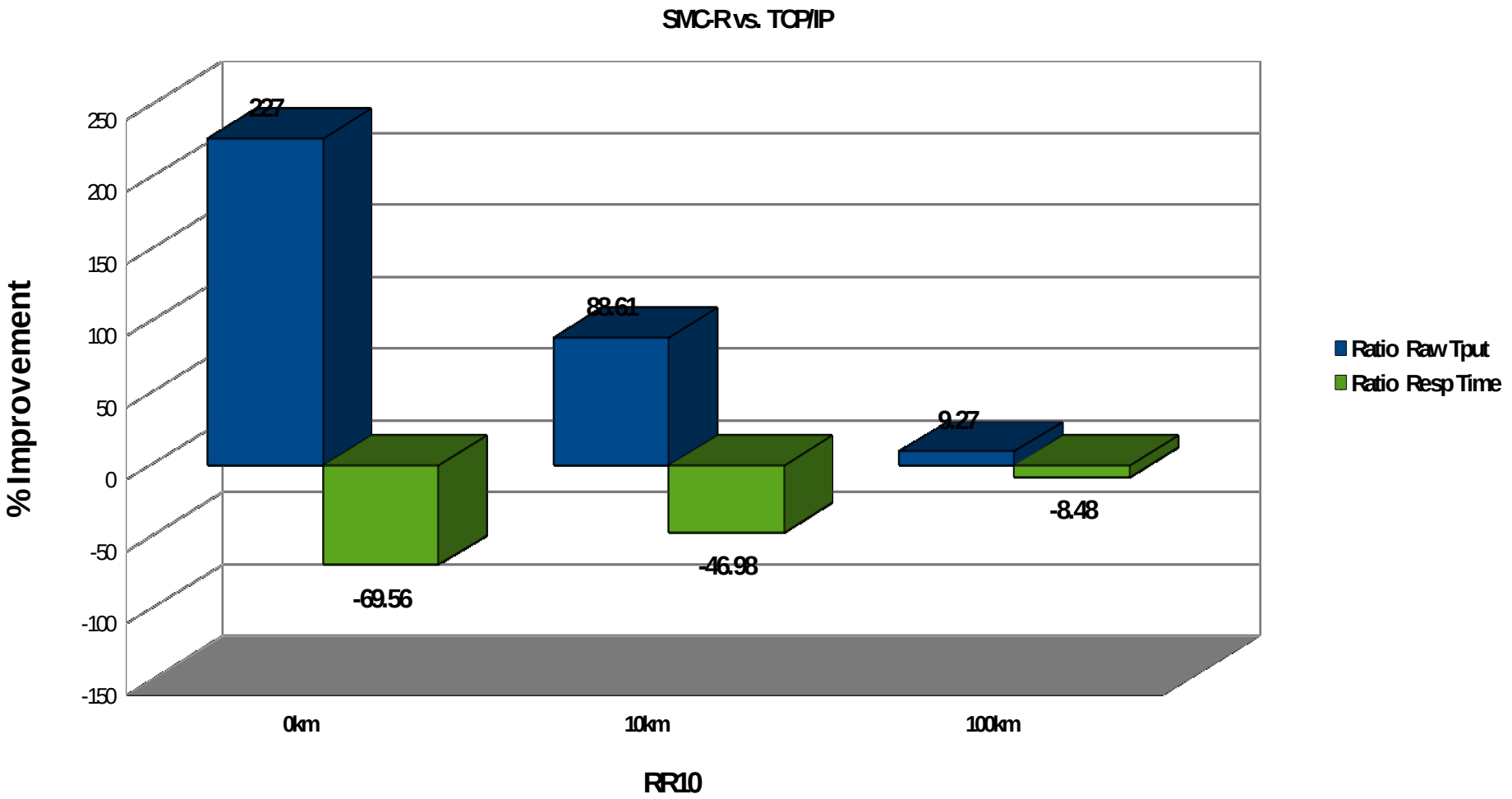
10KM Fiber

100KM Test



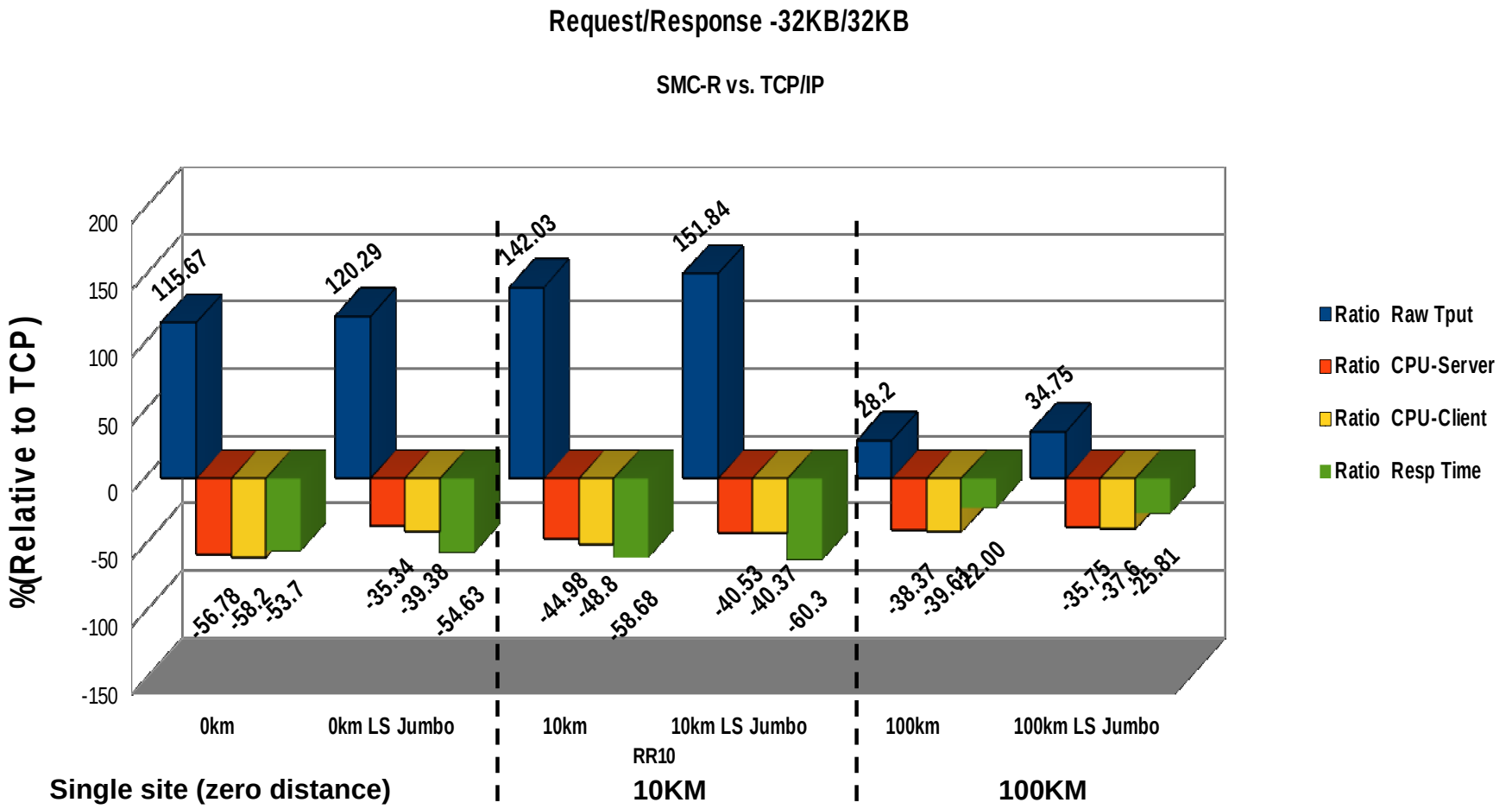
100KM Fiber

SMC-R RoCE performance at distance - Request/Response Pattern (small data)



- Notes:
  - RR10(1K/1K): 10 persistent TCP connections simulating request/response data pattern, client sends 1KB request, server responds with 1KB
  - ***Substantial response time (i.e. latency) improvements at 10KM, benefits drop off at 100km***

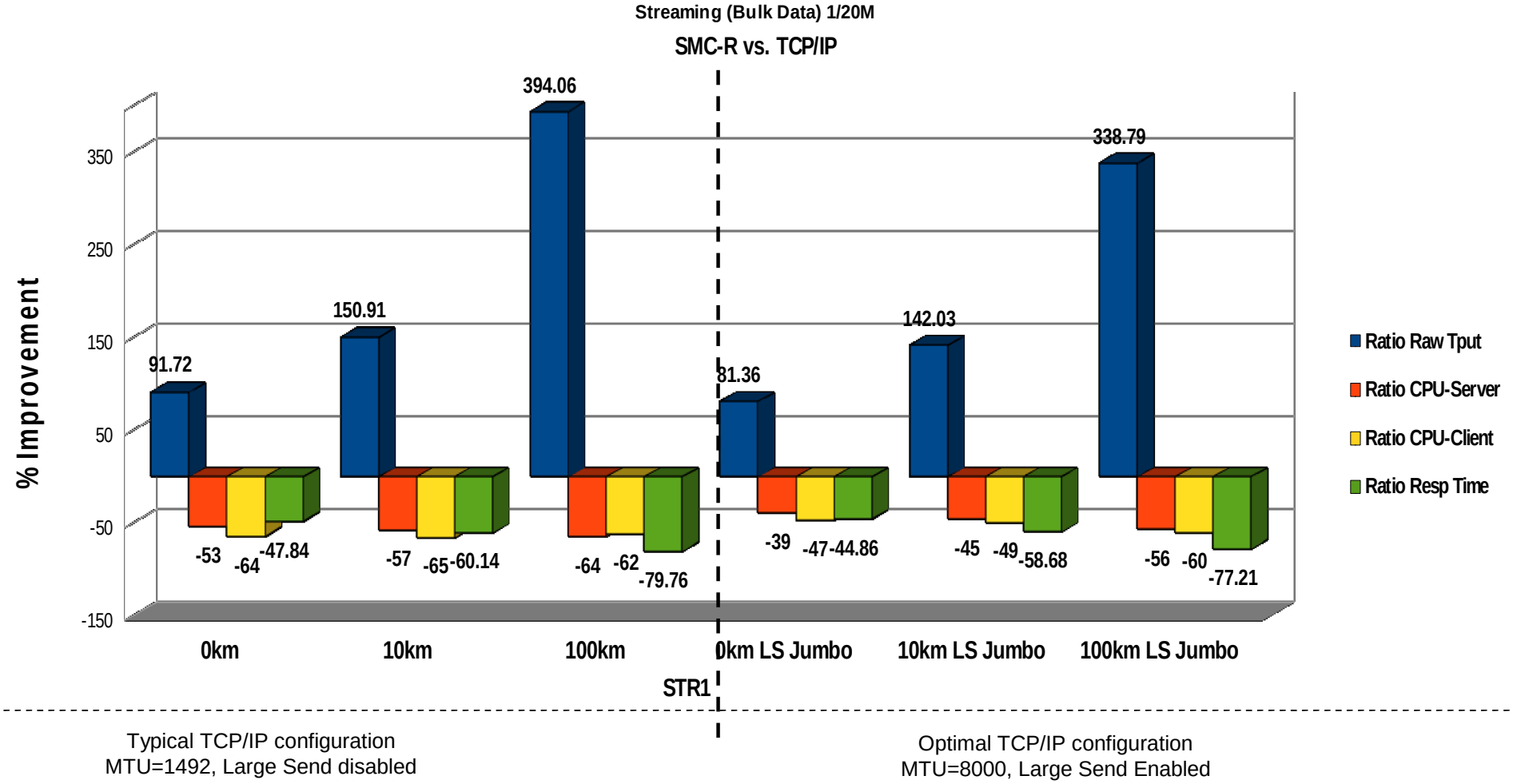
SMC-R RoCE performance at distance – Request/Response Pattern



1) SMC-R vs TCP/IP typical configuration (MTU=1492, Large Send Disabled)
2) SMC-R vs TCP/IP optimal configuration optimal TCP/IP configuration (MTU=8000, Large Send Enabled)

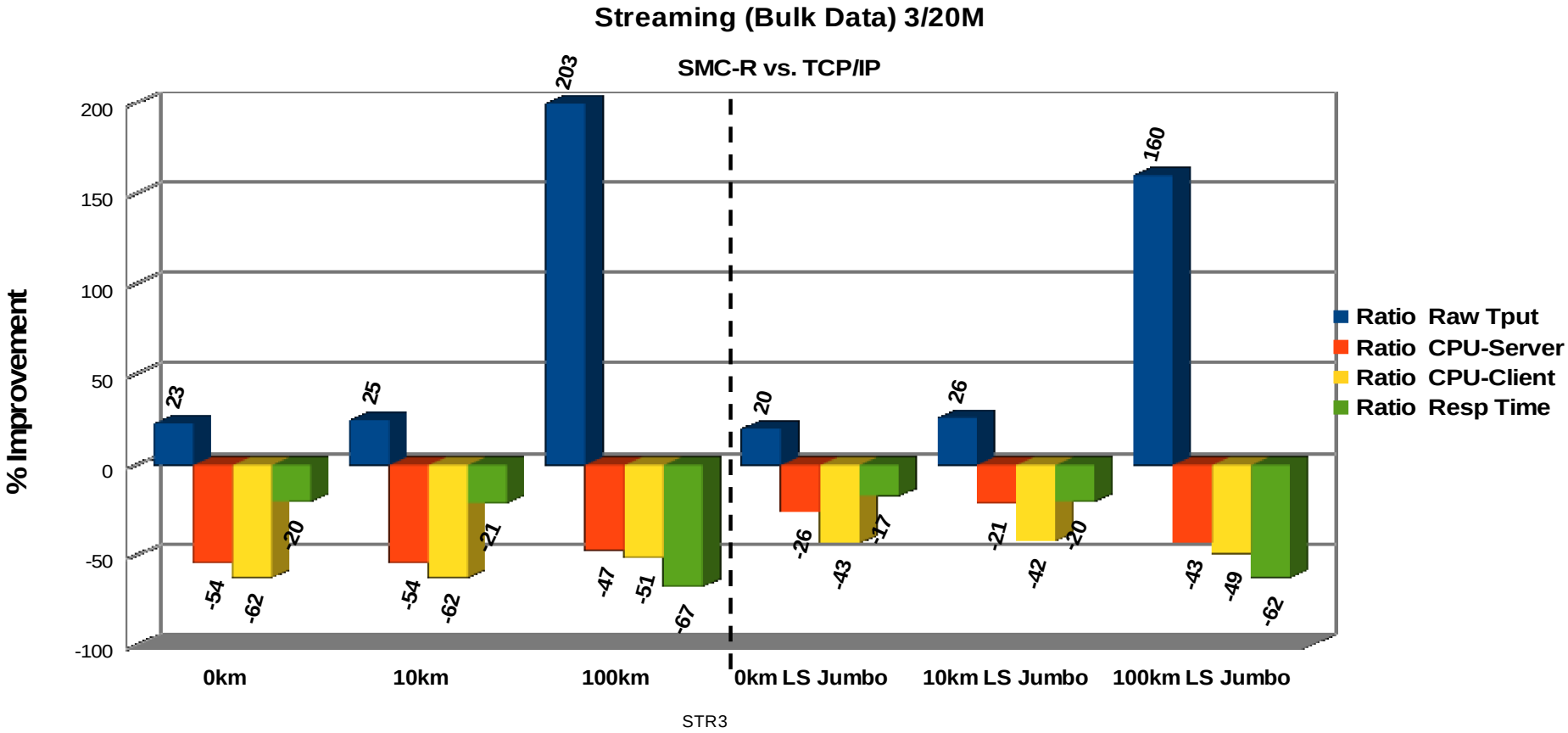
- Notes:
  - RR10(32K/32K): 10 persistent TCP connections simulating request/response data pattern, client sends 32KB request, server responds with 32KB .
  - ***CPU benefits of SMC-R for streaming connections unaffected by distance (and in several cases better at longer distances)***
  - ***Significant response time improvement***

SMC-R RoCE performance at distance – Streaming/Bulk Data (1 session)



- Notes:
  - STR1: Single TCP connection simulating streaming data pattern, client sends 1 byte, server responds with 20MB of data.
  - ***CPU benefits of SMC-R for streaming connections unaffected by distance (and in several cases better at longer distances)***
  - ***Significant throughput improvements at distance (improving overall response time significantly)***

SMC-R RoCE performance at distance – Streaming/Bulk Data (3 sessions)



Typical TCP/IP configuration
MTU=1492, Large Send disabled

Optimal TCP/IP configuration
MTU=8000, Large Send Enabled

- Notes:
  - STR3: Three concurrent TCP connections simulating streaming data pattern, client sends 1 byte, server responds with 20MB of data.
  - ***CPU benefits of SMC-R for streaming connections unaffected by distance (and in several cases better at longer distances)***
  - ***Significant throughput improvements at distance***

Summary of performance benchmarks of SMC-R at distance

- Micro-benchmarks performed at 10km (native ethernet) and 100km (with DWDM) distances
  - At 10km
    - Request/Response workloads (1K/1K payloads): up to 47% lower latency and up to 88% higher throughput than TCP/IP
    - Request/Response workloads (32K/32K payloads): up to 60% lower latency and up to 150% higher throughput than TCP/IP
    - Streaming workloads (20M in one direction): Up to 60% improvement in latency and up to 150% throughput improvement vs TCP/IP
  - At 100km
    - Request/Response workloads (1K/1K payloads): up to 9% lower latency and up to 9% higher throughput than TCP/IP
    - Request/Response workloads (32K/32K payloads): up to 25% lower latency and up to 35% higher throughput than TCP/IP
    - Streaming workloads (20M in one direction): Over 80% improvement in latency and 394% throughput improvement vs TCP/IP (single connection)
  - CPU benefits of SMC-R for larger payloads consistent across all distances

- ***NOTE:*** Based on internal IBM benchmarks using a modeled socket workload in a controlled laboratory environment using micro benchmarks. Your results may vary based on your configuration, workloads and environment.

Summary of performance benchmarks of SMC-R at distance (cont)

- Performance summary
  - Technology viable even at 100km distances with DWDM
  - At 10km: Retain significant latency reduction and increased throughput
  - At 100km: Large savings in latency and significant throughput benefits for larger payloads, modest savings in latency for smaller payloads
  - CPU benefits of SMC-R for larger payloads consistent across all distances

- Use cases for SMC-R at distance
  - TCP Workloads deployed on Parallel Sysplex spanning sites
  - Software based replication (i.e. TCP based) across sites (Disaster Recovery)
    - e.g. InfoSphere Data Replication suite for z/OS
  - File transfers across z/OS systems in different site
    - FTP, Connect:Direct, SFTP, etc.
  - Opportunity: Lower CPU cost for sending/receiving data while boosting throughput and lowering latency