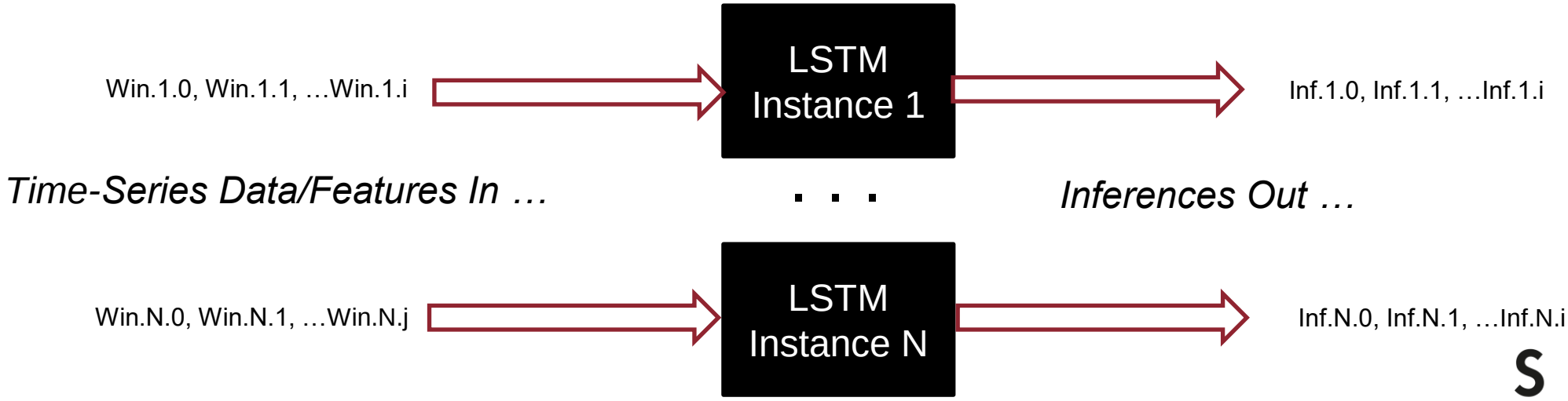
STAC Update
Fast Data & Compute

Jack Gidding
CEO, STAC

jack.gidding@STACresearch.com

STAC-ML Markets (Inference) : Basics

- LSTM models inferring on simulated market data features
- Goal: isolate inference performance of:
  - Inference engine hardware/software
  - Underlying processors, memory, accelerators, etc.
- Metrics:
  - Latency, throughput, error, power efficiency, space efficiency, cost
- Benchmarks allow any level of precision (including mixed-precision)

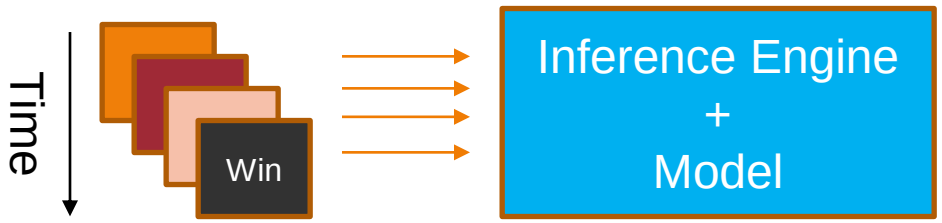


STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

Two benchmark suites

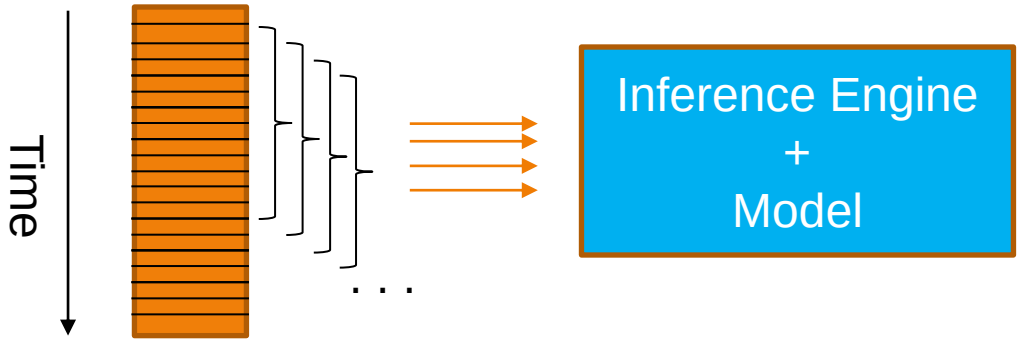
Sumaco

- Operates on fully populated, unique windows of time-series data/features
- Examples:
  - Inference over the recent past in response to an asynchronous event
  - One model may be used to reason about multiple instruments



Tacana

- Operates on sliding windows of a single time-series of data/features
- Example:
  - Inference every tick or bar
- May provide lowest possible tick-to-inference latency



STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

Groq STAC-ML Tacana! (Vault Report)

- Groq was first vendor to publicly report results (Sumaco); Now back with Tacana!

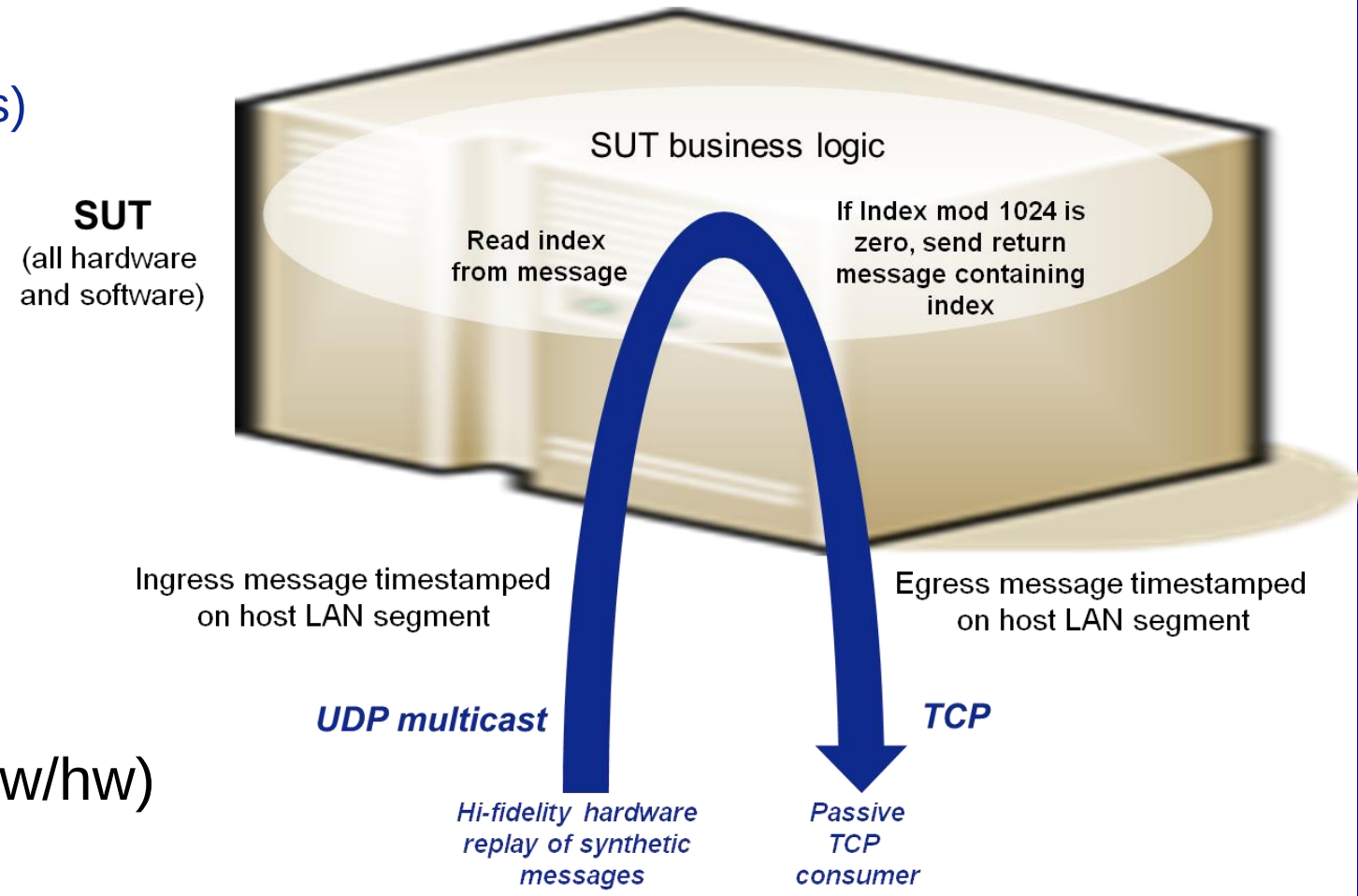
- STAC-ML Tacana Pack for GroqWare™ (Rev A)
- GroqWare™ SDK 0.9.2 devtools and runtime
- C++20 (g++ 11.4.0); Python 3.8.15
- Ubuntu Linux 22.04 LTS
- GroqNode™ GN1-B8C-ES:
  - 1 x GroqCard™ 1 Accelerators (GC1-010B)
  - 2 x AMD EPYC™ 7413 24-core CPUs @ 2650 MHz
  - 16 slots x 64GiB DDR4 - 1024GiB Total
- Report is available to STAC-ML & Trade Flow Subscribers

www.STACresearch.com/GROQ231106

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-T0

- Tick-to-trade pattern
  - UDP in (507B or 68B frames x 3 rates)
  - TCP out (122B frames)
- Isolates network I/O latency
  - Latency that cannot be squeezed out of business logic
  - Does not co-mingle I/O latency with market-specific logic
- Extremely high accuracy
- Works with any trading platform (sw/hw)
- Key metric: Actionable Latency
  - STAC-T0.ACTIONABLE.LAT: $t_{FO} - t_{IND}$



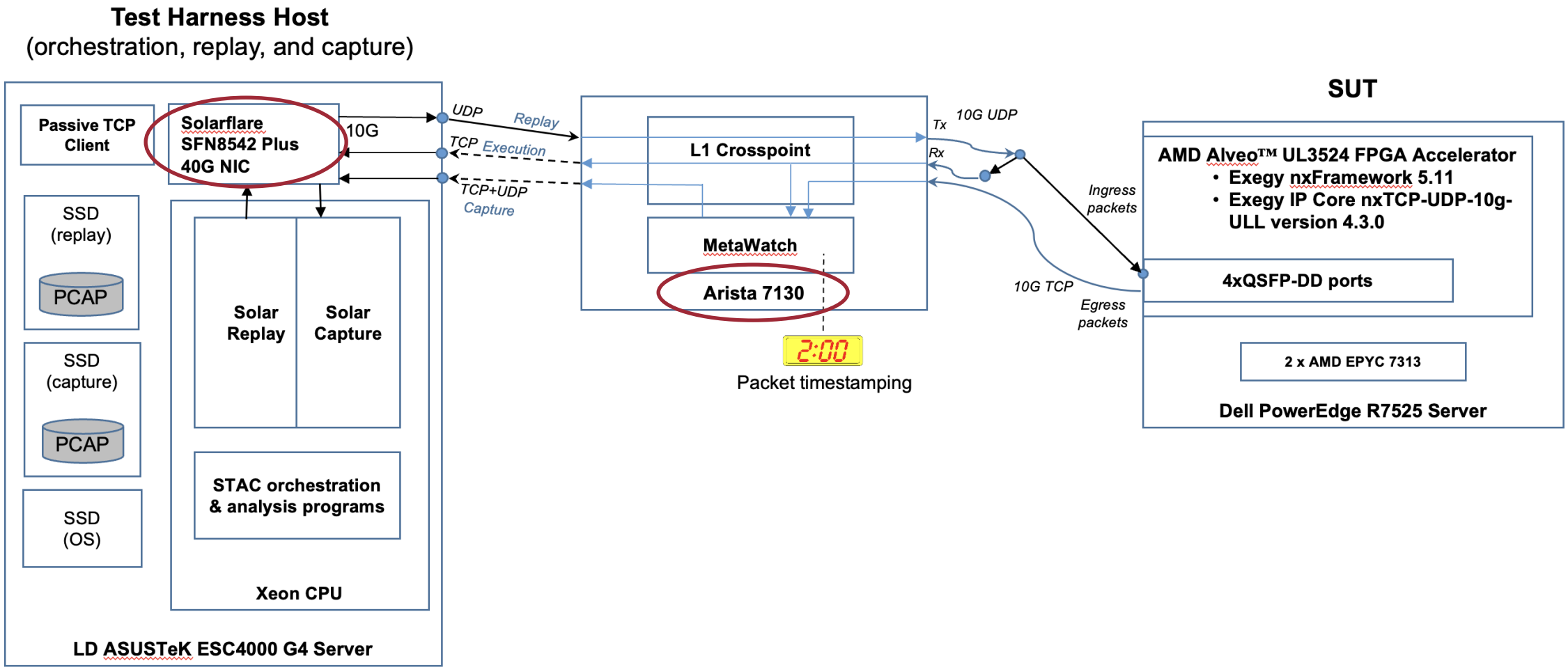
STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

Breaking News: New solution from AMD and Exegy (AMD240422)

- AMD Alveo™ UL3524 FPGA Accelerator
- Exegy tick-to-trade reference design, including:
  - Exegy nxFramework 5.11
  - Exegy IP Core nxTCP-UDP-10g-ULL version 4.3.0
- Dell PowerEdge R7525

Report coming …

Test setup



STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

Results highlights

- New records for all message sizes

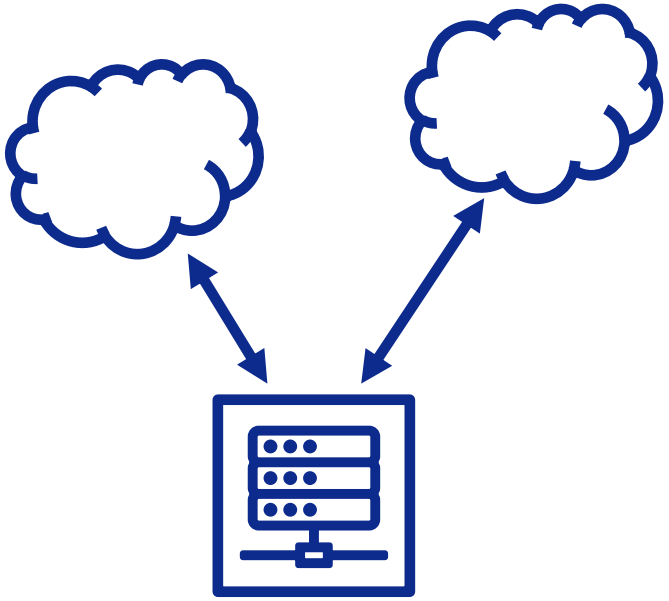
Actionable latency:

- For 507-byte frames at all ingress rates
  - Minimum of 13.9 nanos
    (STAC-T0.β1.*.A.ACTIONABLE.MIN)
  - Maxes lower than the previous records by 39-42%
    (STAC-T0.β1.*.A.ACTIONABLE.MAX)
- For 68-byte frames at MEDRATE & HIGHRATE
  - Minimum of 14.1 nanos
    (STAC-T0.β1.[.B.ACTIONABLE.MIN)
  - Maxes lower than the previous records by 43-48%
    (STAC-T0.β1.[MEDRATE | HIGHRATE].B.ACTIONABLE.MAX)

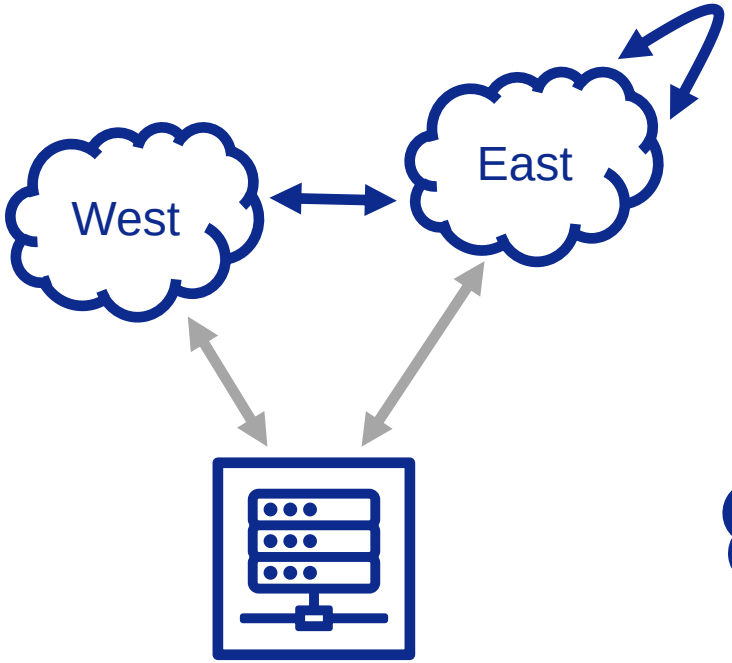
STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-N2

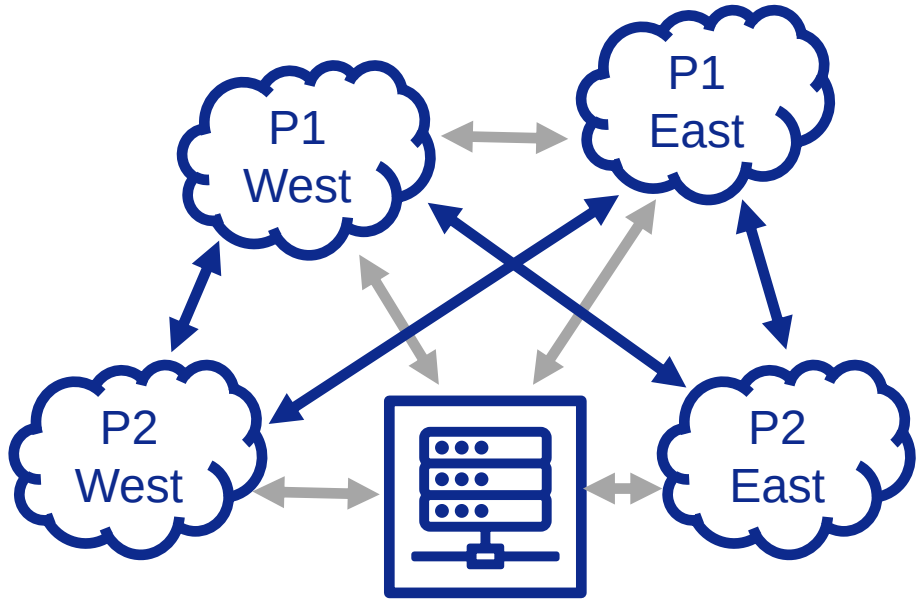
- STAC-N2 is designed to measure the latency and throughput of communication paths between data centers, both on-prem to cloud and cloud-cloud.



On-prem to Cloud



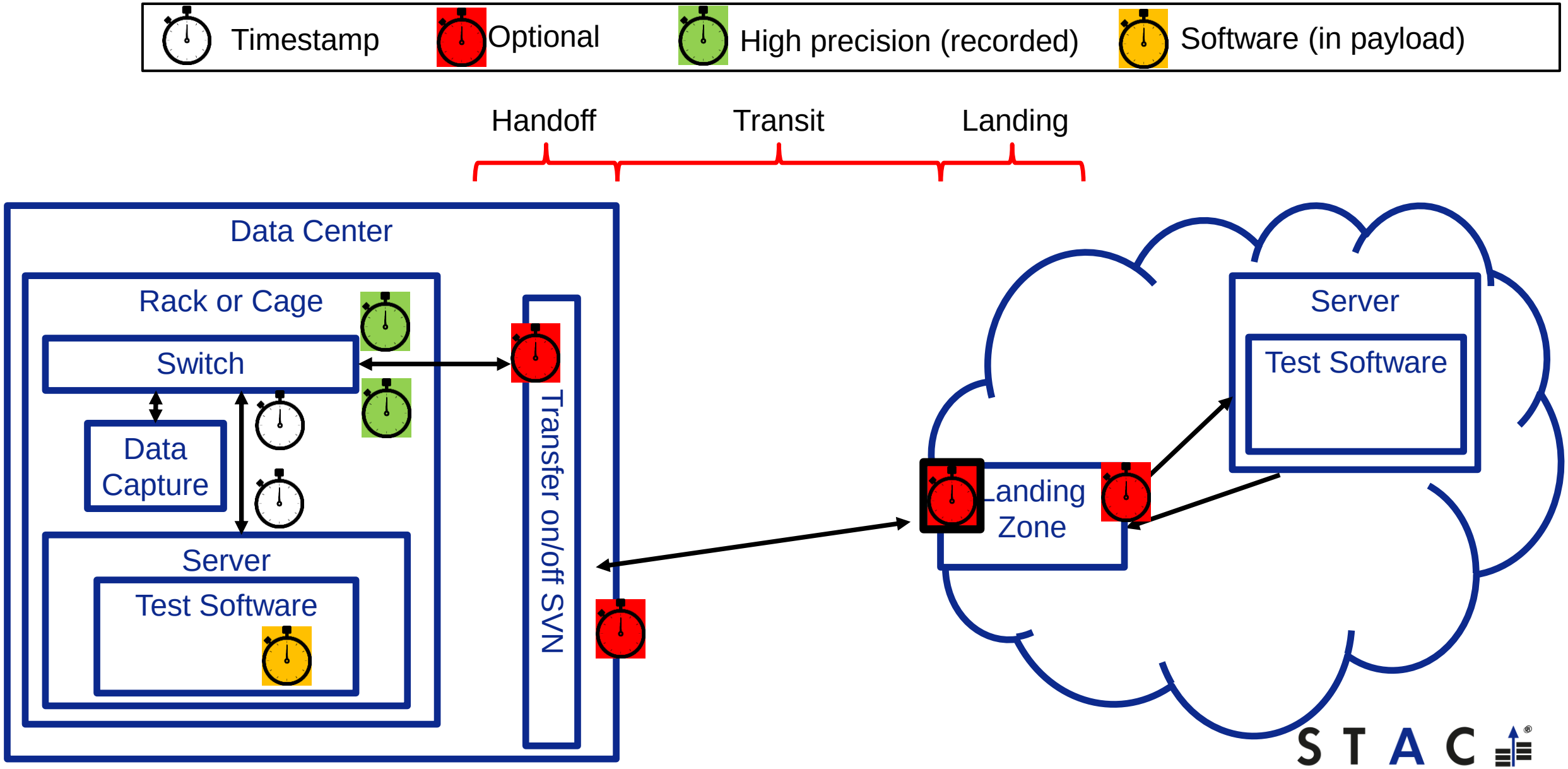
Cloud Region to Cloud Region



Cloud Provider to Cloud Provider

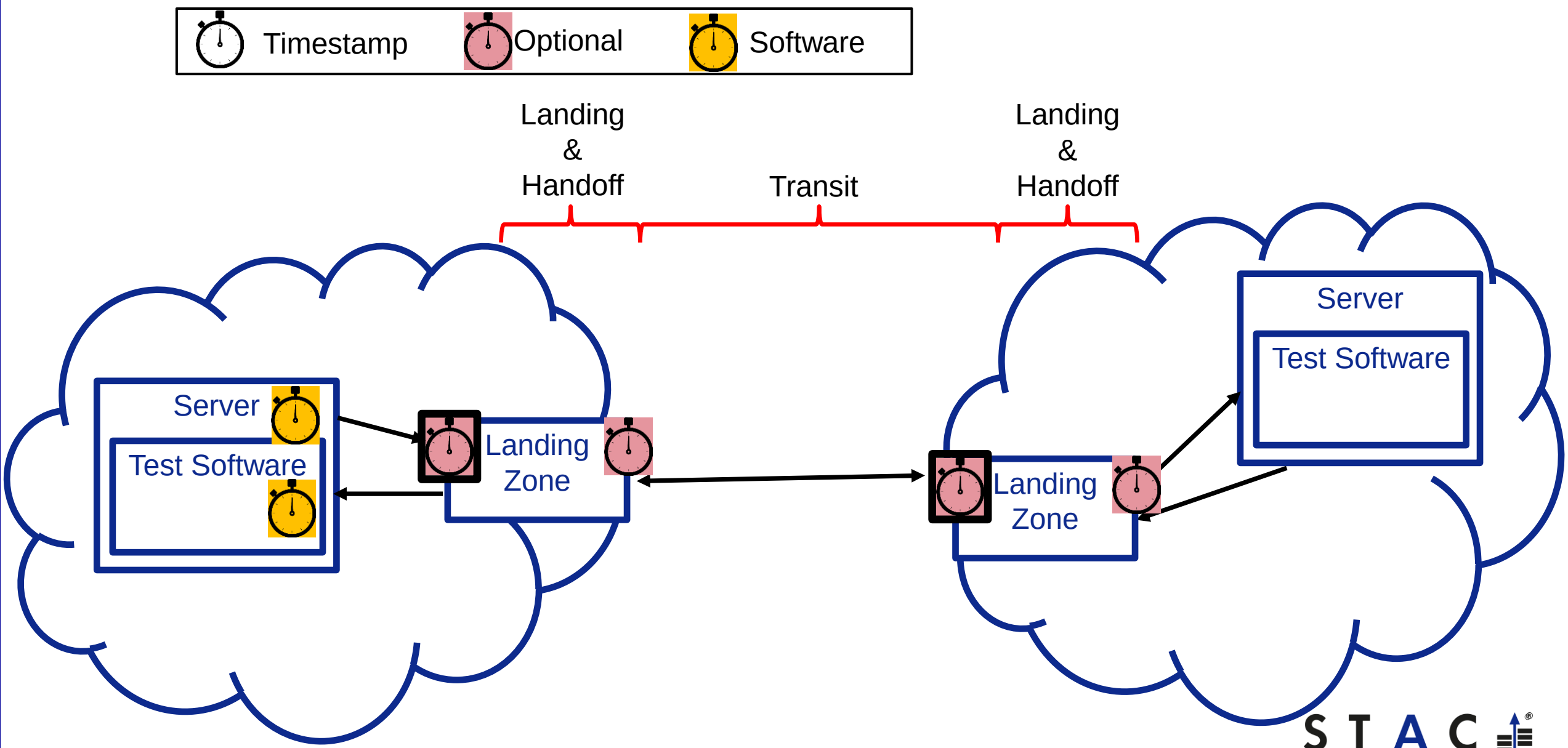
STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

On-prem-to-cloud



STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

Cloud-to-cloud

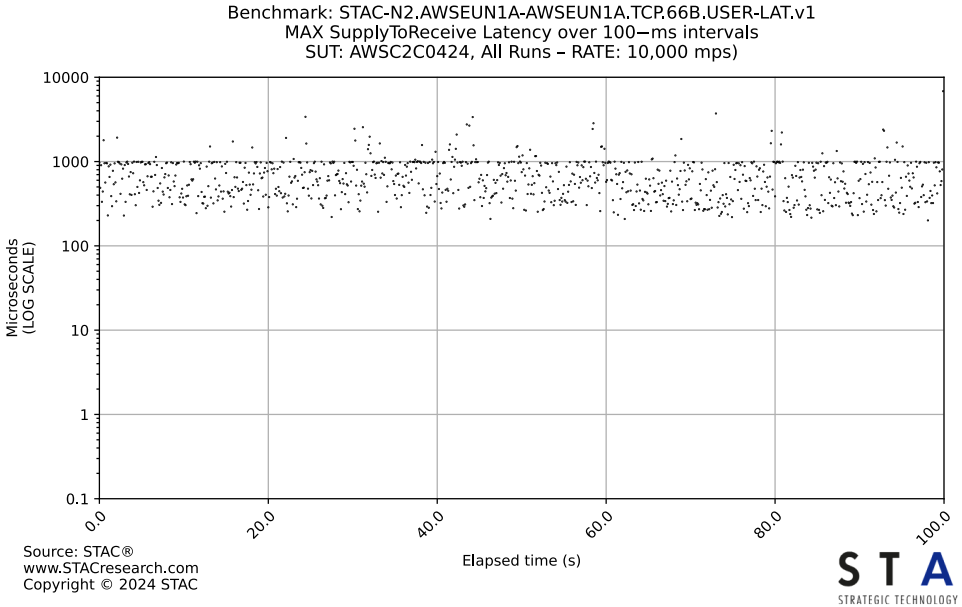


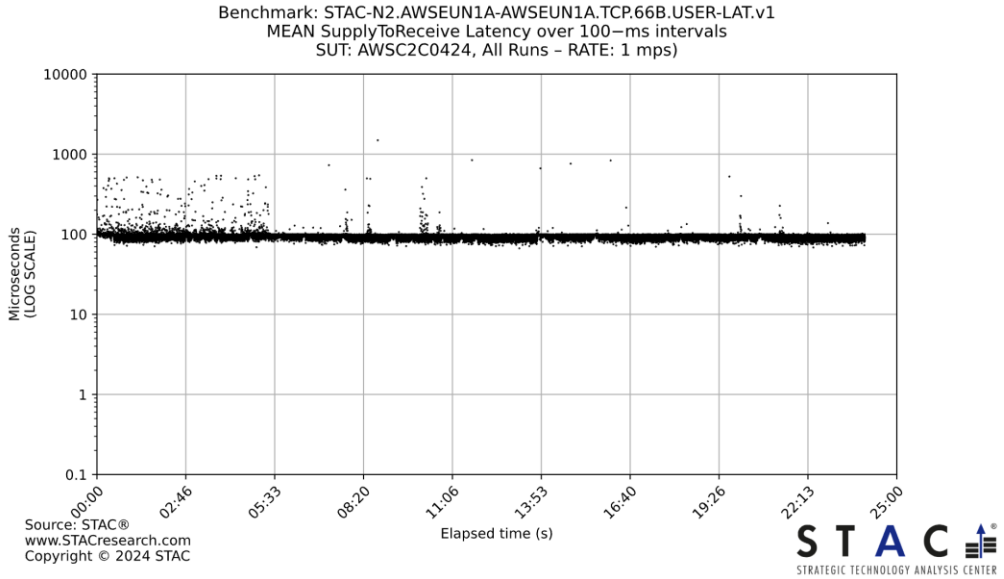
STAC
STRATEGIC TECHNOLOGY ANALYSIS CENTER

Status

council@STACresearch.com

- 3 packet sizes (66/528/1000B) and 3 data rates (10k/50k/max rate) for 1M msgs + 24-hour duration test for each packet size at 1 mps
- Reliable and/or unreliable protocols
- STAC-N2 Benchmark Spec Rev D is latest
- Test Harness is available (in development)





STAC-N2.AWSEUN1A-AWSEUN1A.TCP.66B.USER-LAT.v1
SUT: AWSC2C0424

Latency statistics (us)

Rate (msgs/sec)	Minimum	Median	Mean	Maximum	Std Dev
1	67	93	106.0	12005	178.6
10,000	34	111	162.0	11676	271.2
50,000	35	434	622.0	22345	888.5

Source: STAC
www.STACresearch.com
Copyright © 2024 STAC

STAC
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER