Hewlett Packard
Enterprise

EARLY EXPERIENCE IN SUPPORTING OPENSHMEM ON SLINGSHOT 11

Naveen Namashivayam, Bob Cernohous, Nathan Wichmann, Mark Pagel

Cray User Group (CUG), 2022

TECHNICAL DATA RIGHTS

All materials contained in, attached to, or referenced by this document that are marked Cray Confidential, HPE Proprietary & Confidential, or with a similar restrictive legend may not be disclosed in any form without the advance written permission of Hewlett Packard Enterprise (HPE). These data are submitted with limited rights under Government Contract No. B626589 and Lease Agreement 4000167127. These data may be reproduced and used by the Government with the express limitation that they will not, without written permission of HPE, be used for purposes of manufacture nor disclosed outside the Government.

This notice shall be marked on any reproduction of these data, in whole or in part.

COPYRIGHT AND TRADEMARK ACKNOWLEDGEMENTS

FORWARD LOOKING STATEMENTS

This presentation may contain forward-looking statements that involve risks, uncertainties and assumptions. If the risks or uncertainties ever materialize or the assumptions prove incorrect, the results of Hewlett Packard Enterprise Company and its consolidated subsidiaries ("Hewlett Packard Enterprise") may differ materially from those expressed or implied by such forward-looking statements and assumptions. All statements other than statements of historical fact are statements that could be deemed forward-looking statements, including but not limited to any statements regarding the expected benefits and costs of the transaction contemplated by this presentation; the expected timing of the completion of the transaction; the ability of HPE, its subsidiaries and Cray to complete the transaction considering the various conditions to the transaction, some of which are outside the parties' control, including those conditions related to regulatory approvals; projections of revenue, margins, expenses, net earnings, net earnings per share, cash flows, or other financial items; any statements concerning the expected development, performance, market share or competitive performance relating to products or services; any statements regarding current or future macroeconomic trends or events and the impact of those trends and events on Hewlett Packard Enterprise and its financial performance; any statements of expectation or belief; and any statements of assumptions underlying any of the foregoing. Risks, uncertainties and assumptions include the possibility that expected benefits of the transaction described in this presentation may not materialize as expected; that the transaction may not be timely completed, if at all; that, prior to the completion of the transaction, Cray's business may not perform as expected due to transaction-related uncertainty or other factors; that the parties are unable to successfully implement integration strategies; the need to address the many challenges facing Hewlett Packard Enterprise's businesses; the competitive pressures faced by Hewlett Packard Enterprise's businesses; risks associated with executing Hewlett Packard Enterprise's strategy; the impact of macroeconomic and geopolitical trends and events; the development and transition of new products and services and the enhancement of existing products and services to meet customer needs and respond to emerging technological trends; and other risks that are described in our Fiscal Year 2018 Annual Report on Form 10-K, and that are otherwise described or updated from time to time in Hewlett Packard Enterprise's other filings with the Securities and Exchange Commission, including but not limited to our subsequent Quarterly Reports on Form 10-Q. Hewlett Packard Enterprise assumes no obligation and does not intend to update these forward-looking statements.

BACKGROUND

- PGAS (Partitioned Global Address Space)
  - Style of programming
  - Employs light-weight one-sided asynchronous communication primitives
  - Exploits NIC RDMA features – remote read/write without target process involvement
  - Examples: Language (UPC, Coarrays in Fotran), Library (Global Arrays, GASPI, OpenSHMEM)

- OpenSHMEM
  - Community standard one-sided API specification
  - Provides PGAS (Partitioned Global Address Space) style of programming
  - Used to achieve scalable performance on applications with irregular communication patterns
    - Integer Sort, Sparse solvers, and FFT
    - Explored and used as an alternate for message passing style programming

- Cray OpenSHMEMX
  - HPE proprietary SW Library implementation
  - Available as part of HPE Cray Programming Environment SW stack
  - Actively developed !!

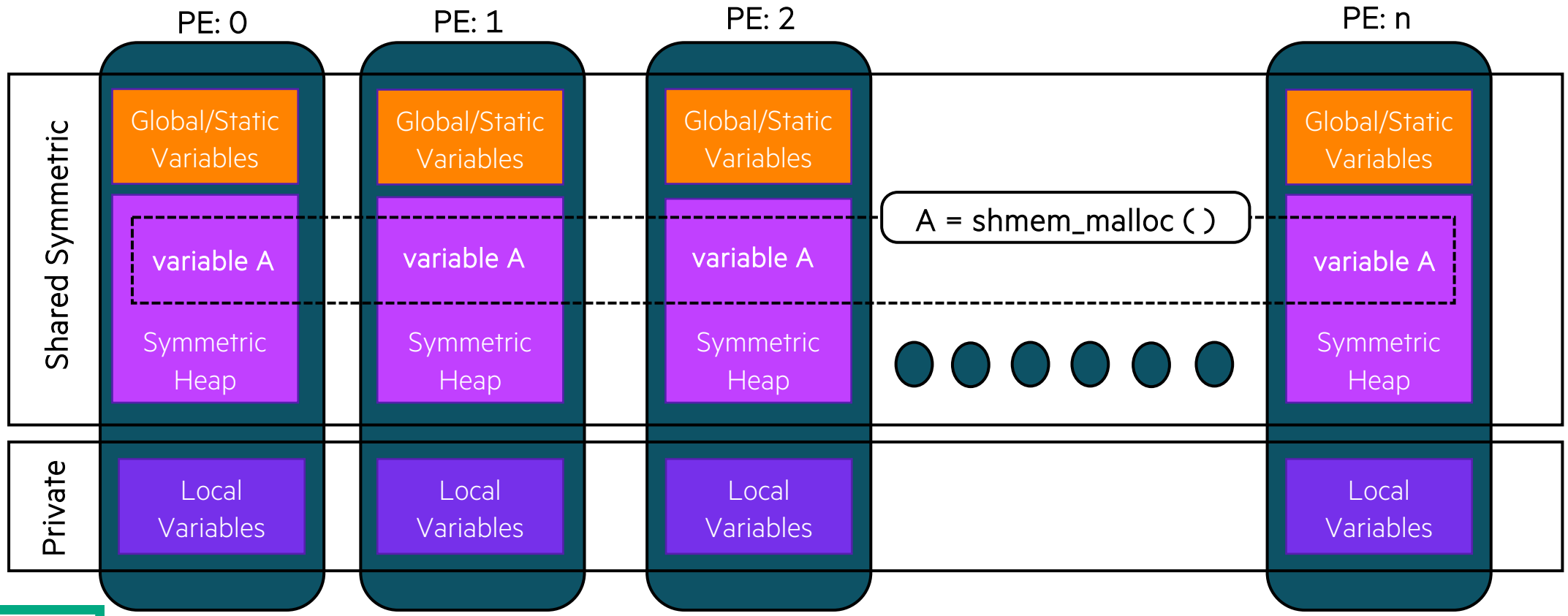
MOTIVATION

- HPE Slingshot interconnect
  - HPE Rosetta switch
  - Industry standard NIC (Slingshot 10)
    – Mellanox ConnectX-5 100 Gb/s Ethernet NICs
  - HPE Slingshot NIC (Slingshot 11)
    – HPE proprietary NIC
    – Planned to power the three announced US exascale systems

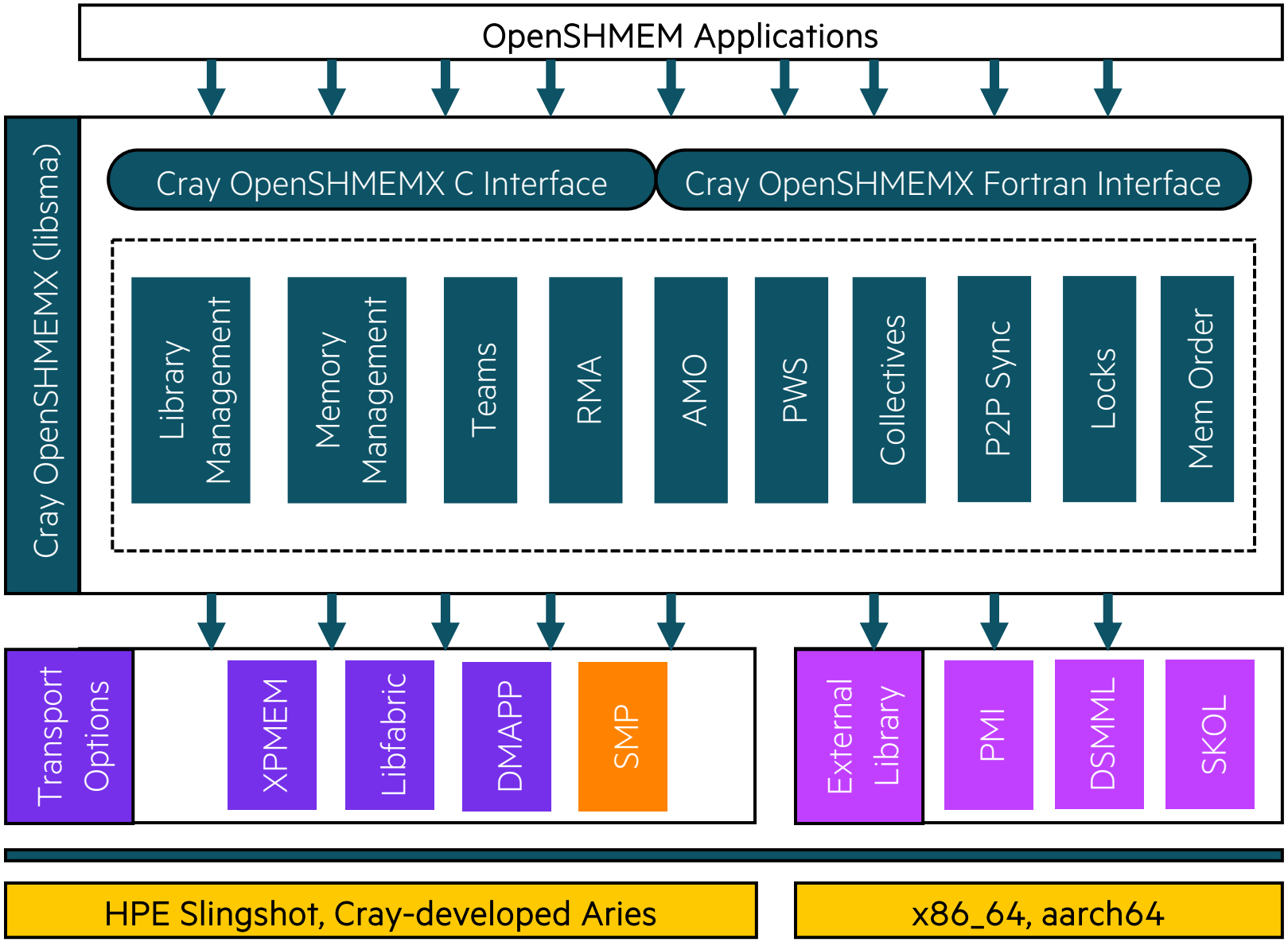
- Motivation
  - Use OpenSHMEM as an exemplar to evaluate Slingshot 11 NIC
  - Introduce and report various new Slingshot 11 features
    – Specifically, that impacts the performance of OpenSHMEM operations
  - Propose different non-standard implementation-specific features in Cray OpenSHMEMX
    – Exploit best performance from Slingshot 11

OPENSHMEM PMODEL OVERVIEW

- SPMD-style of programming
- Memory model - (1) private and (2) shared symmetric heap
- Move data between symmetric data objects using RMA, AMO, or collective operations
  - Example: Remote write RMA operation - *shmem_put (src, dst, nelems, target-PE)*



CRAY OPENSHMEMX OVERVIEW



SLINGSHOT 11 NIC FEATURES

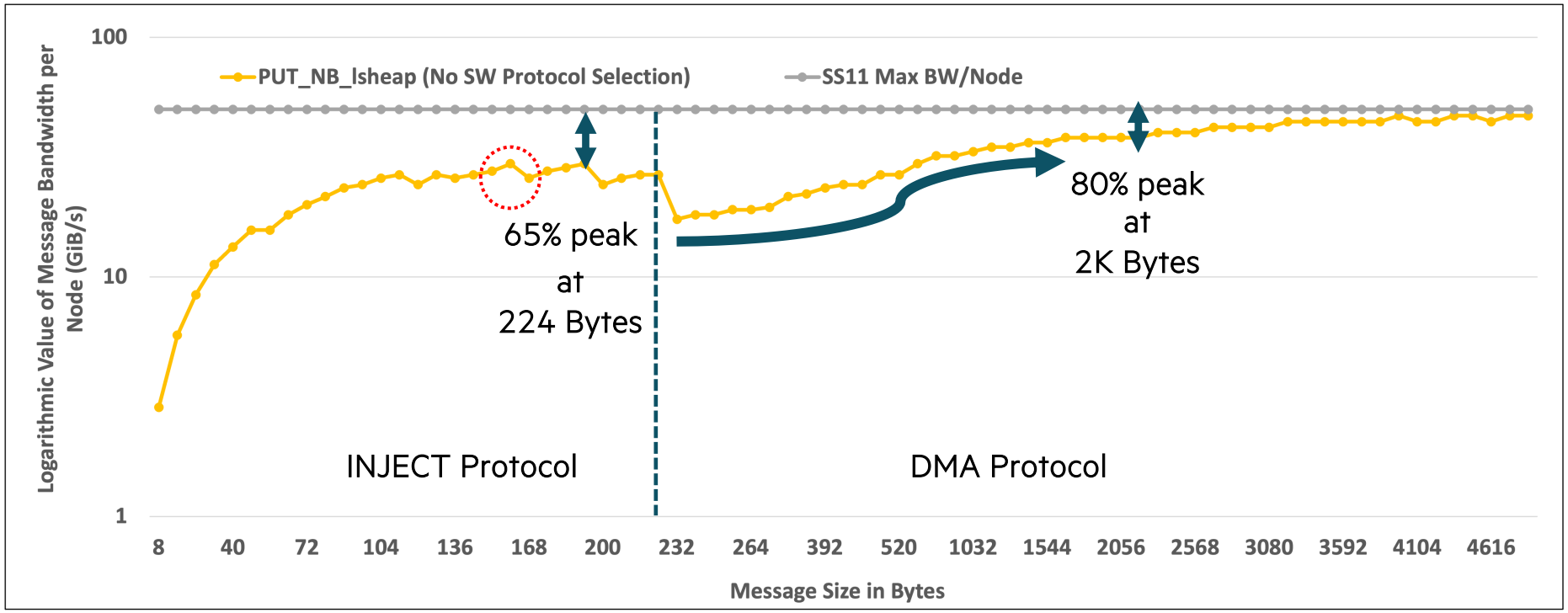
- Overview of Slingshot 11 features – impacting the performance of OpenSHMEM operations

- RMA transfer protocol
  - Determines how data is consumed by Slingshot 11 NIC
- Event completion semantics
  - Early completion notification vs. remote target memory completion notification
- Bundling communication events
  - Transfer single event vs. multiple events together
- AMO reliability
  - Single transmission vs. multiple transmission of AMO events

SLINGSHOT 11 NIC FEATURE – RMA PROTOCOL ANALYSIS: 1

- INJECT – message packed with packet header
- DMA – message sent separate (as single or multiple chunks) from the header
- *shmem_put_nbi (const void * source, void * target, size_t size, int target-PE-index)*
- Data size and source buffer type are important factors
  - Target buffer – always shared symmetric variable
  - Source buffer – (1) symmetric heap, or (2) local heap/stack

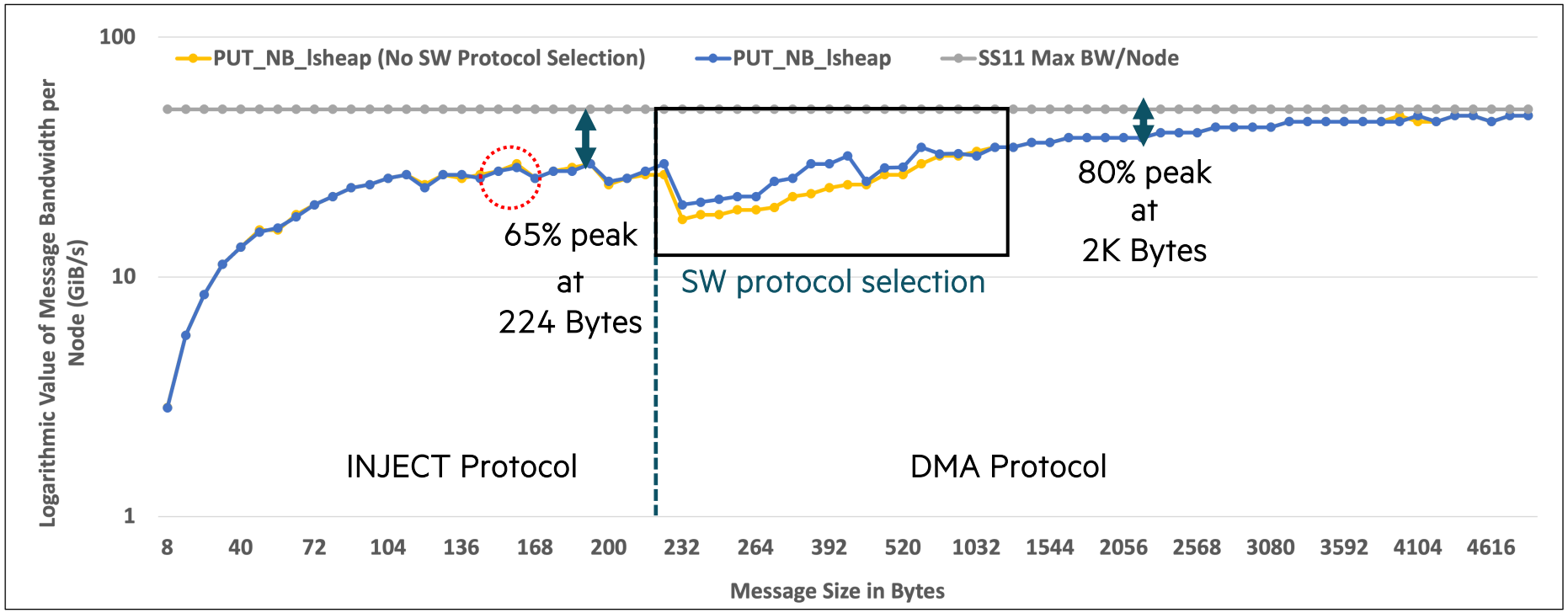
16 Nodes, 128 PPN
Dual-NIC AMD Milan
HPE Random Access MB
NBI PUT BW Analysis



SLINGSHOT 11 NIC FEATURE – RMA PROTOCOL ANALYSIS: 1

- INJECT – message packed with packet header
- DMA – message sent separate (as single or multiple chunks) from the header
- *shmem_put_nbi (const void * source, void * target, size_t size, int target-PE-index)*
- Data size and source buffer type are important factors
  - Target buffer – always shared symmetric variable
  - Source buffer – (1) symmetric heap, or (2) local heap/stack

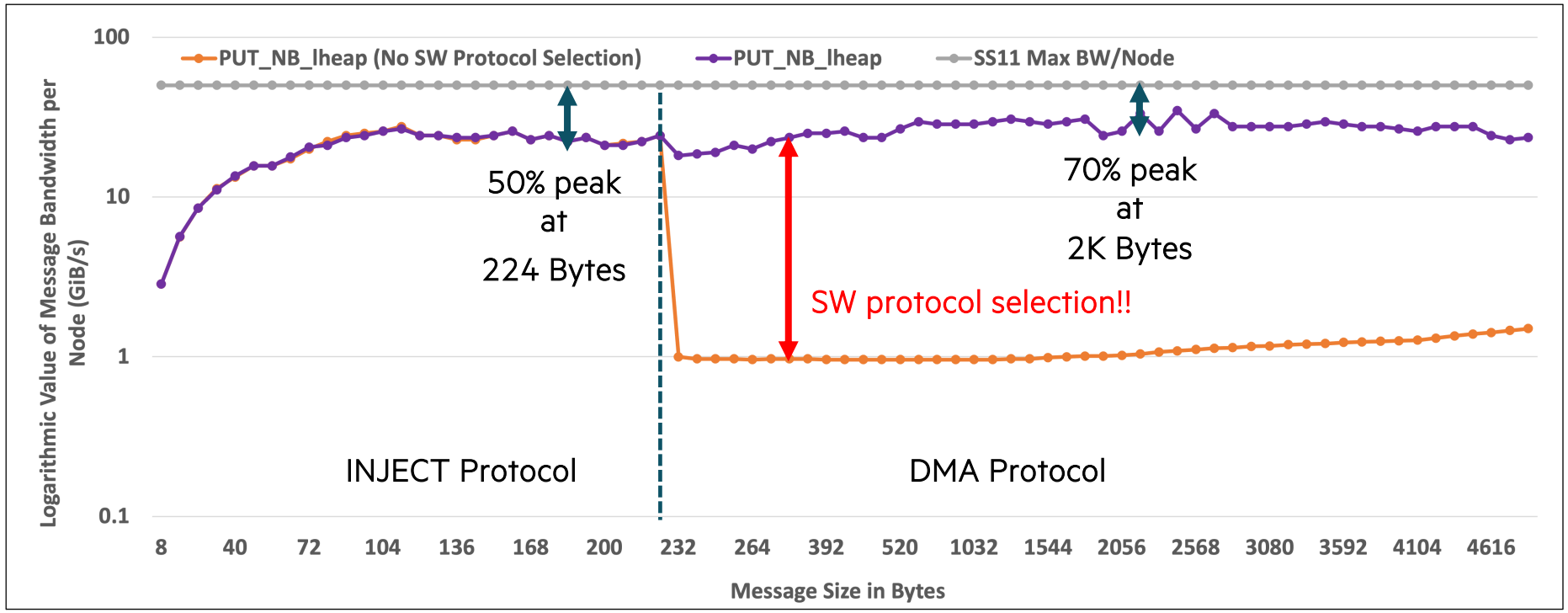
16 Nodes, 128 PPN
Dual-NIC AMD Milan
HPE Random Access MB
NBI PUT BW Analysis



SLINGSHOT 11 NIC FEATURE – RMA PROTOCOL ANALYSIS: 3

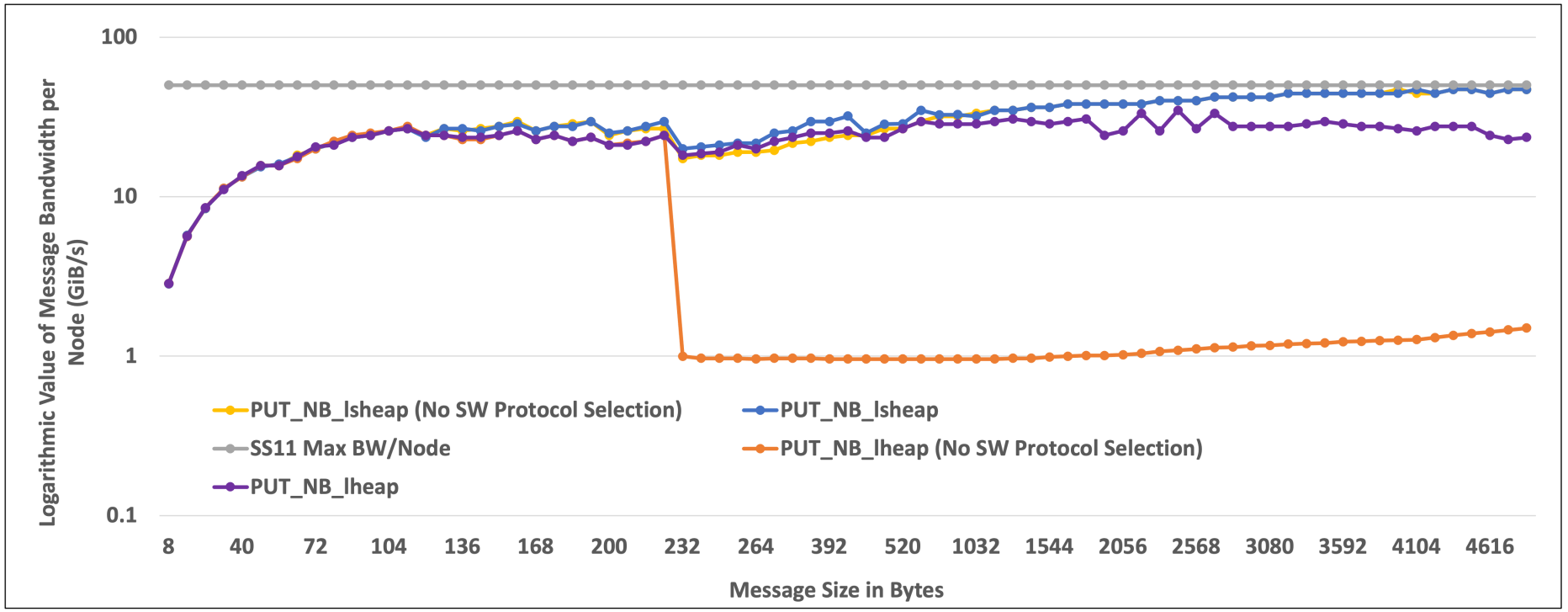
- INJECT – message packed with packet header
- DMA – message sent separate (as single or multiple chunks) from the header
- *shmem_put_nbi (const void * source, void * target, size_t size, int target-PE-index)*
- Data size and source buffer type are important factors
  - Target buffer – always shared symmetric variable
  - Source buffer – (1) symmetric heap, or (2) local heap/stack

16 Nodes, 128 PPN
Dual-NIC AMD Milan
HPE Random Access MB
NBI PUT BW Analysis



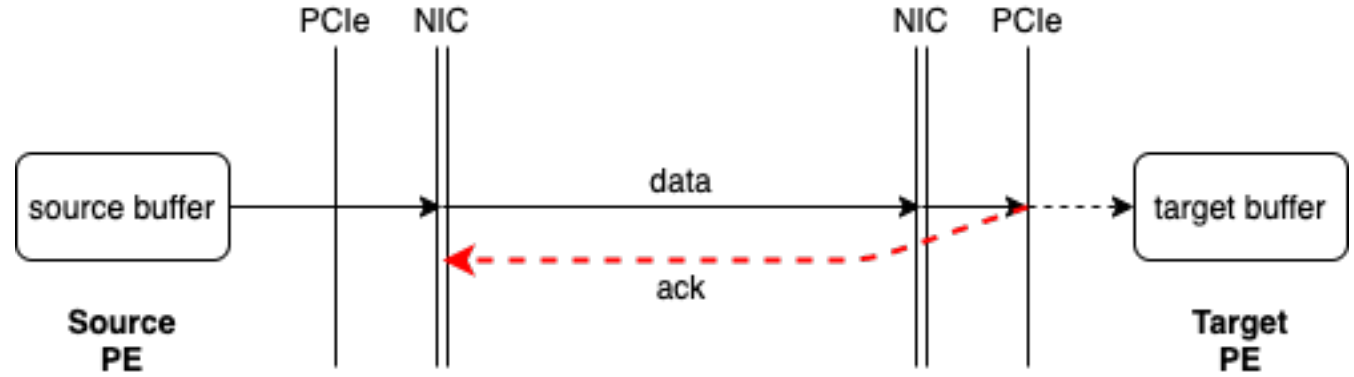
SLINGSHOT 11 NIC FEATURE – RMA PROTOCOL ANALYSIS: 4

- Overall Inference
  - Use symmetric heap-based source buffers
  - Using local heap/stack-based source buffers have performance impact
  - For small-message transfers – data size is critical to achieve peak performance
  - Transitioning between protocol has performance impact – use SW protocol selection module (tuning in-progress)

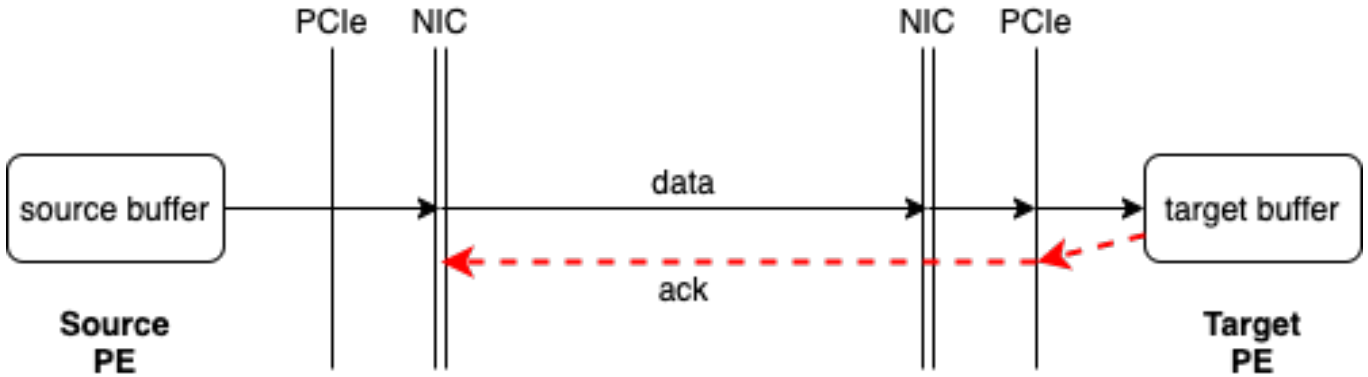


SLINGSHOT 11 NIC FEATURE – EVENT COMPLETION SEMANTICS

- Determines when an OpenSHMEM operation is completed
- *Types of completions supported by Slingshot 11 - (1) Local completion, and (2) Remote target completion



Local Completion



Remote target completion

*Different from the completion protocol supported by the libfabric middleware. This slide represents the true completions supported by the Slingshot 11 NIC.

SLINGSHOT 11 NIC FEATURE – SUPPORTED COMPLETIONS

- OpenSHMEM supported completions
  - **shmem_quiet ( )** – Guarantees remote target completion of all previous posted operations
  - **shmem_fence ( )** – Guarantees ordering of operations when delivered on the remote target memory

	Local completion	Remote target completion	Event ordering
shmem_put (blocking remote write)	Return from operation	Return from **shmem_quiet**	Return from **shmem_fence**
shmem_put_nbi (non-blocking remote write)	Not supported	Return from **shmem_quiet**	Return from **shmem_fence**

- No option to provide local completions for non-blocking operations
  - How to exploit the local completion semantics offered by Slingshot 11 NIC ?
  - Cray OpenSHMEMX specific extensions

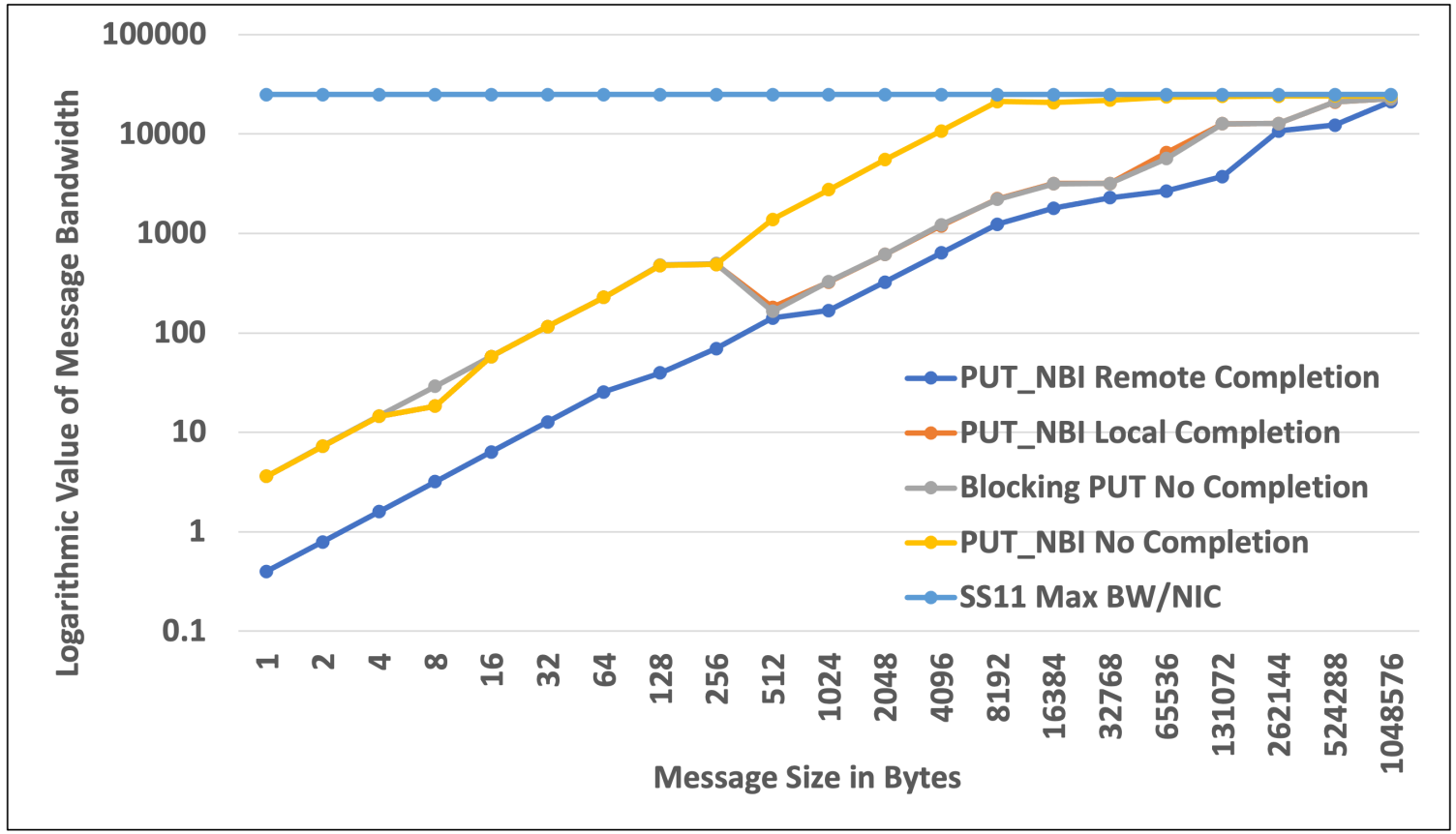
void **shmemx_local_complete** (void);
void **shmemx_ctx_local_complete** (shmem_ctx_t ctx);

SLINGSHOT 11 NIC FEATURE – EVENT COMPLETION ANALYSIS

- Comparing using blocking and non-blocking PUT operations with different completion semantics

- How to use non-blocking PUT operations with local completion semantics ?
  - Group a bunch of non-blocking PUTs
  - Communication-Computation overlap

2 Nodes, 1 PPN
Dual-NIC AMD Milan
OSU MB Suite
PUT BW Analysis

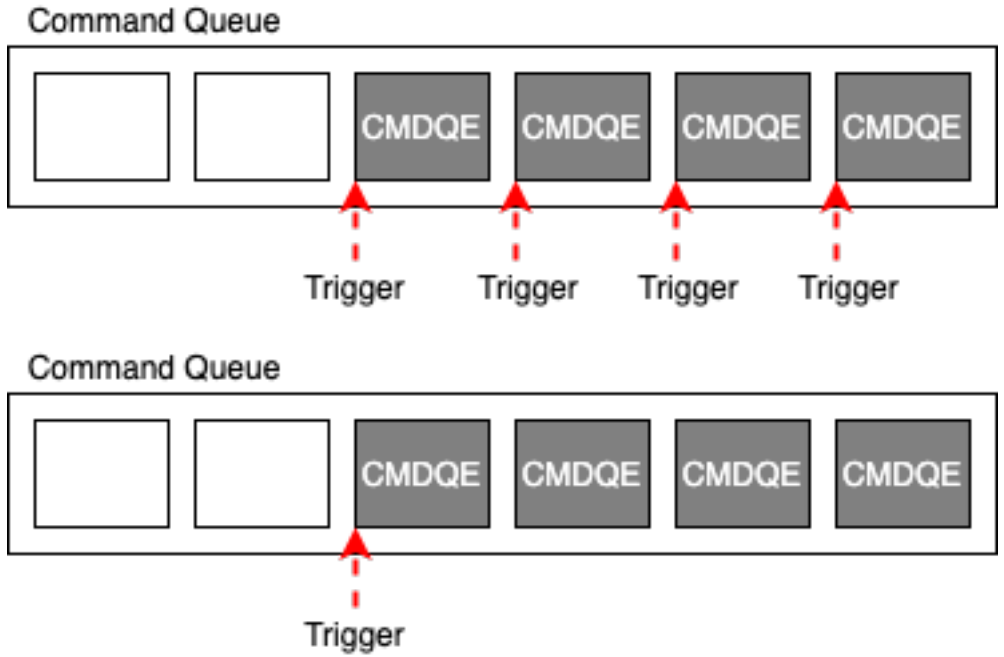


SLINGSHOT 11 NIC FEATURE – BUNDLING EVENTS

- How to post event in Slingshot 11 ?
  - Prepare and enqueue message packet into Command Queue as CMDQE
  - Trigger the enqueued Command Queue Entries (CMDQE)
  - NIC executes CMDQE as FIFO

- How to trigger CMDQE ?
  - Trigger every event
  - Trigger group of events

- Bundling
  - Group CMDQE and trigger as single operation



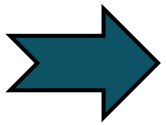
SLINGSHOT 11 NIC FEATURE – BUNDLING SUPPORT

- No option to bundle OpenSHMEM operations
- Cray OpenSHMEMX-specific extensions – SESSIONs

- OpenSHMEM Sessions
  - Epoch in an application
  - Users provide hints for the epoch
  - Hints allow better NIC resource management
  - Similar to ***#pragma*** directives in C language

void **shmemx_ctx_session_start** (IN shmem_ctx_t ctx, IN int options);
void **shmemx_ctx_session_stop** (IN shmem_ctx_t ctx);

```
for (int i = 0; i < n; i++) {
        shmem_put_nbi (SHMEM_CTX_DEFAULT,
                       src + off[i], dst + off[i],
                       nelems, i);
}
shmem_quiet();
```



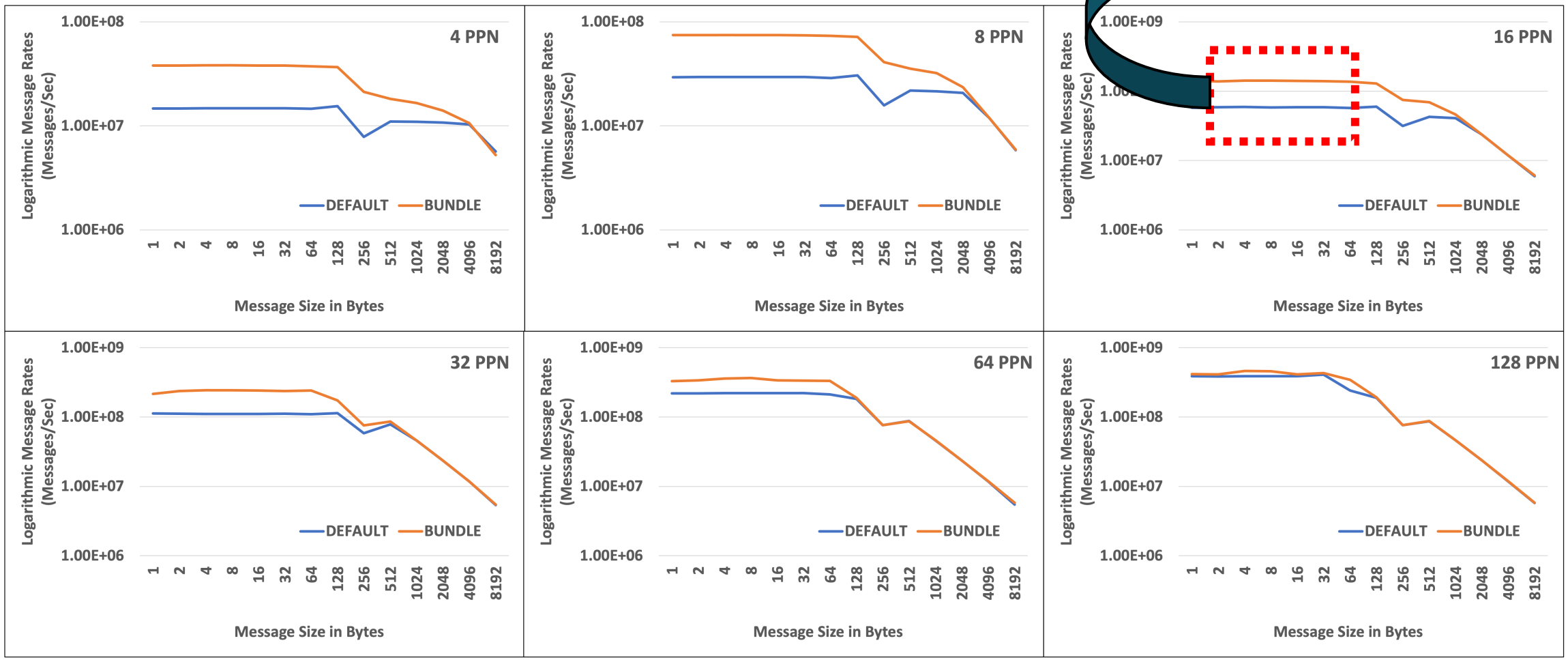
```
shmemx_ctx_session_start (SHMEM_CTX_DEFAULT,
SHMEM_SESSION_BUNDLE | SHMEM_SESSION_OP_PUT);

for (int i = 0; i < n; i++) {
        shmem_put_nbi (SHMEM_CTX_DEFAULT,
                       src + off[i], dst + off[i],
                       nelems, i);
}

shmemx_ctx_session_stop (SHMEM_CTX_DEFAULT);
shmem_quiet();
```

SLINGSHOT 11 NIC FEATURE – BUNDLING ANALYSIS

- OSU Non-blocking Bundling Analysis – 2 Nodes Different PPN



OVERALL INFERENCE

- Data size and type of data buffer are critical for performance
  - Small-size messages in INJECT protocol category is data size sensitive
  - Large-size messages are always pushed to use DMA protocol
  - SW protocol selection module
    - Critical for source buffers in global or local heap memory
    - Useful for medium size messages with source buffers in symmetric heap memory
- Local completions can help to achieve better computation/communication overlap
- Bundling operations help use-cases with low PPN count
  - 16 PPN case shows 2.5X performance improvement
- Other Slingshot 11 specific analysis are available in the paper

CONCLUSION

- Introduced and explored Slingshot 11 features that impacts OpenSHMEM performance

- Discussed on this presentation
  - RMA Protocol selection
  - Event completion semantics
  - Bundling operations

- Proposed new Cray OpenSHMEMX-specific extensions
  - Exploit Slingshot 11 specific features
  - Bundling OpenSHMEM Sessions
  - Fine-grain Completions
  - Single-transmit AMOs

- Future analysis
  - Explore tuning OpenSHMEM collectives and multithreading for Slingshot 11 NIC

Thank you !