Hewlett Packard
Enterprise

EARLY EXPERIENCE IN SUPPORTING ONE-SIDED COMMUNICATIONS ON SLINGSHOT 11

Naveen Ravi

FoRMA'22

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.

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 – HPE SLINGSHOT

- HPE Slingshot interconnect
  - Discrete PCIe NIC and switch
  - Brings together the best of Ethernet and HPC Interconnects
  - HPE Rosetta – HPE proprietary switch
  - HPE Slingshot NIC (**Slingshot 11**)
    - HPE proprietary NIC
    - Used on ORNL's Frontier exascale system

MOTIVATION

- Introduce and report on various Slingshot 11 features – impacting one-sided communications
- Provide implementation options for MPI RMA

SLINGSHOT 11 NIC FEATURES

- Overview of Slingshot 11 features – impacting the performance of one-sided communication operations

- RMA transfer protocol
  - Determines how data is consumed by Slingshot 11 NIC
  - Inject vs. DMA

- Event completion semantics
  - Early completion notification vs. remote target memory completion notification

- Bundling communication events
  - Transfer single event vs. multiple events together

- Triggered communication operations
  - Deferred execution events

SLINGSHOT 11 NIC FEATURE – RMA PROTOCOLS

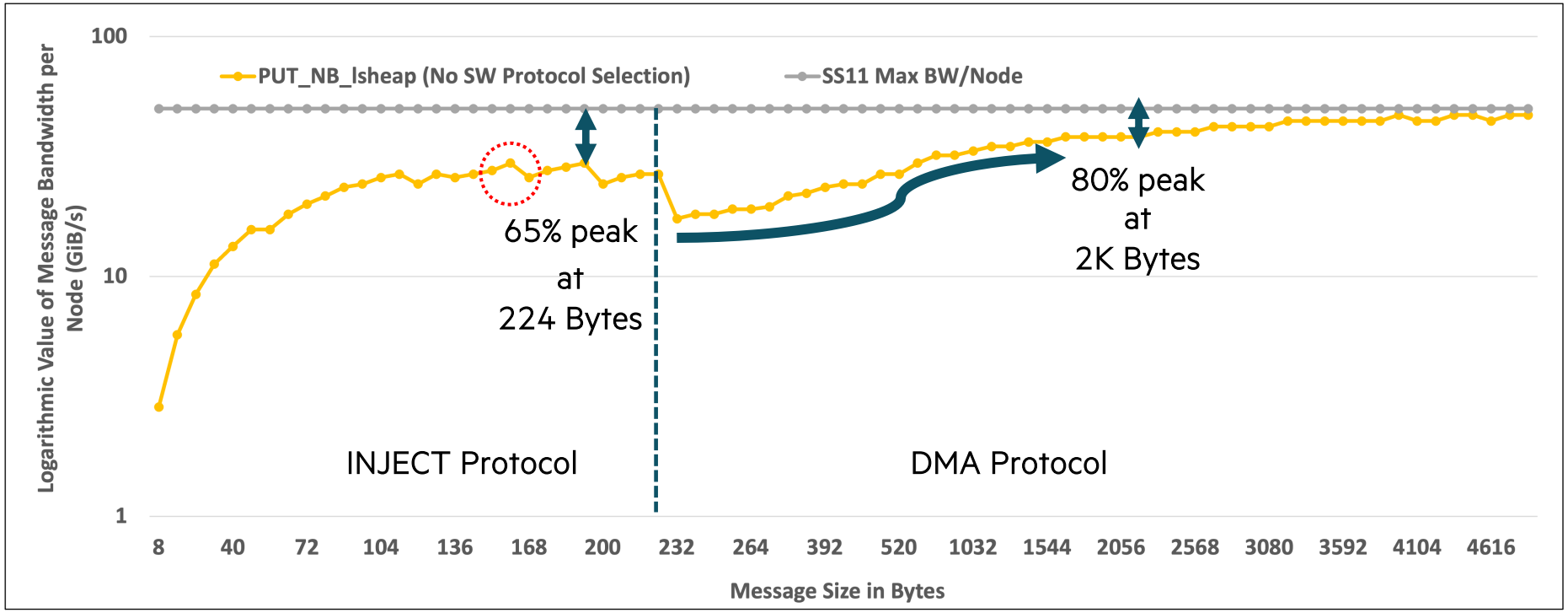
- INJECT
  - Low-latency immediate data commands
  - Message packed with packet header
  - Fixed size limitation

- DMA
  - Message sent separate from the header
  - Message sent as a single or multiple chunks
  - Typically used for large message buffer transfers

RMA PROTOCOL ANALYSIS – DATA SIZE

- Performance impacting factors
  - Data-size – determines RMA protocol selection
  - Source buffer type – determines the need for memory registration

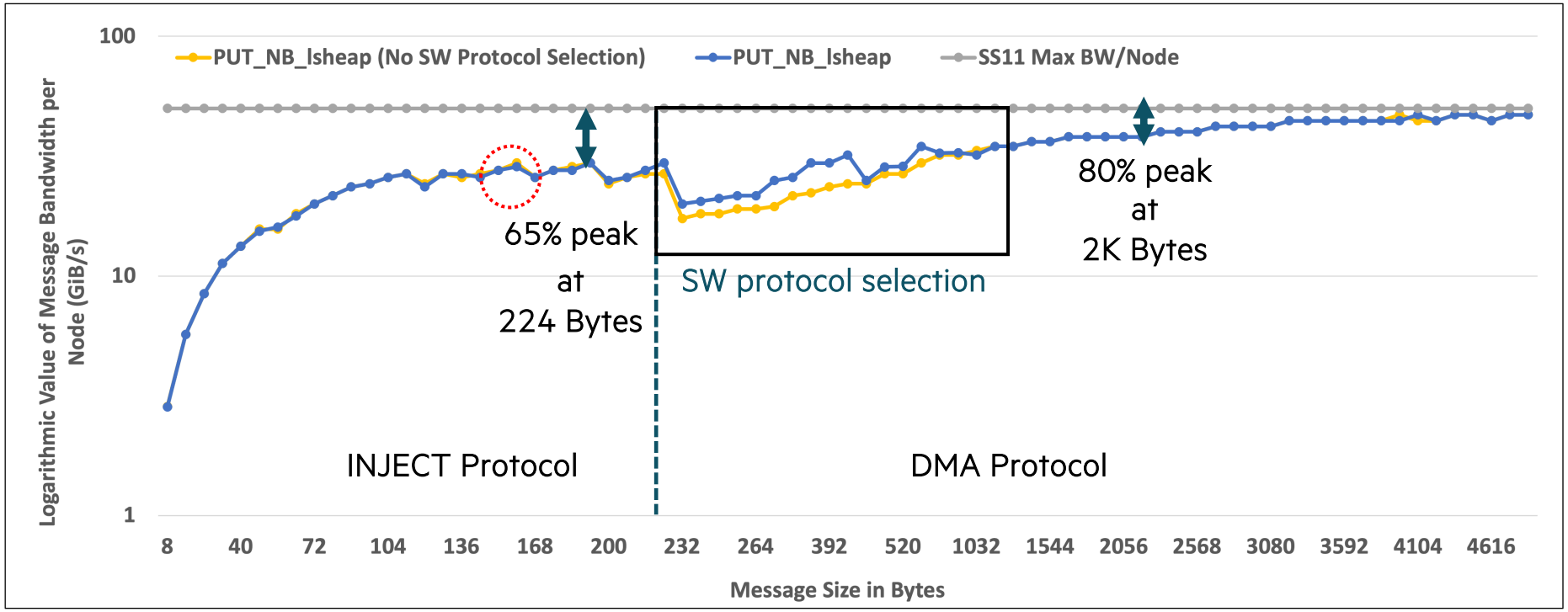
16 Nodes, 128 PPN
Dual-NIC AMD Milan
HPE Random Access MB
NBI PUT BW Analysis



RMA PROTOCOL ANALYSIS – SW PROTOCOL SELECTION

- Performance impacting factors
  - Data-size – determines RMA protocol selection
  - Source buffer type – determines the need for memory registration

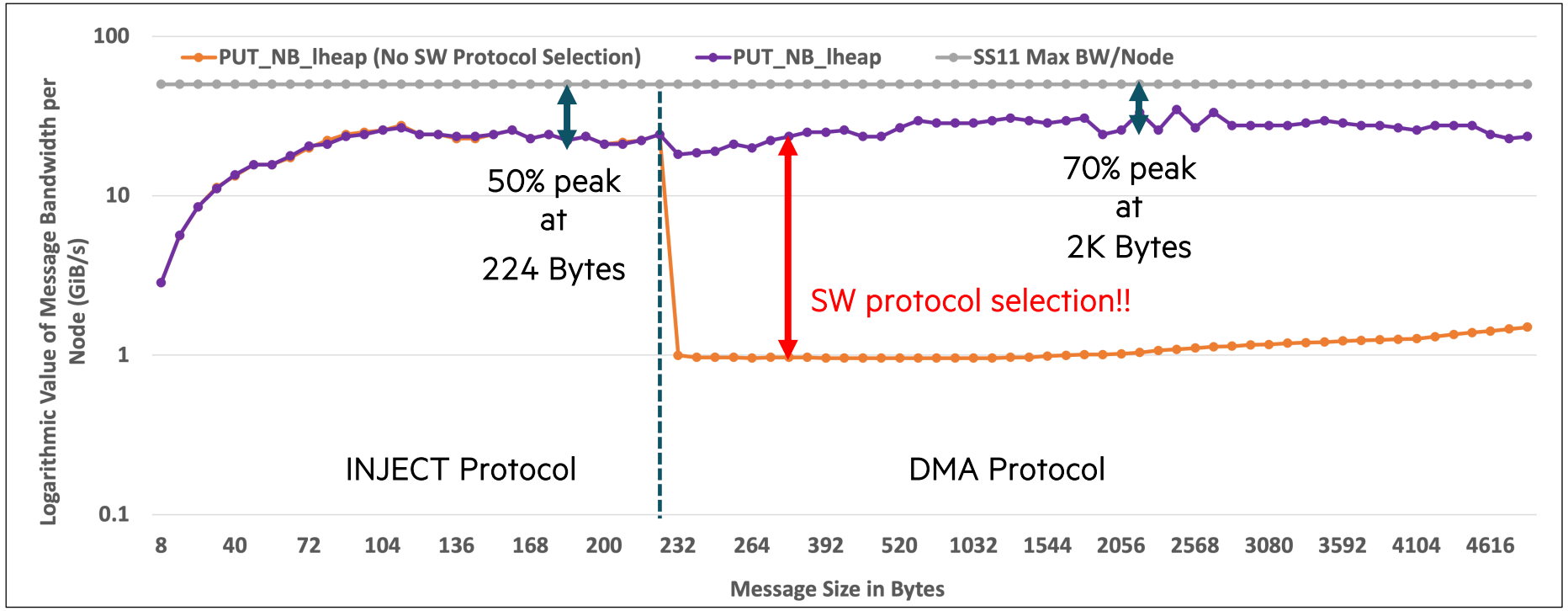
16 Nodes, 128 PPN
Dual-NIC AMD Milan
HPE Random Access MB
NBI PUT BW Analysis



RMA PROTOCOL ANALYSIS – SOURCE BUFFER TYPE

- Impact of source buffer type
  - Target buffer – always a registered memory space
  - 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



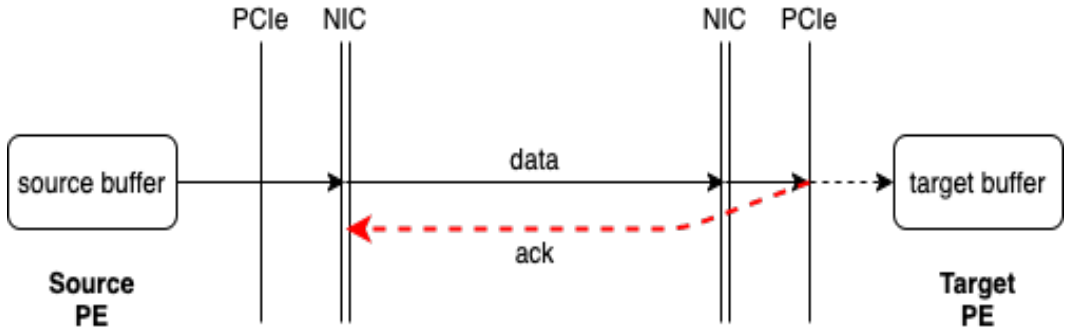
RMA PROTOCOL ANALYSIS

- Overall Inference
  - Use registered memory region as source buffers
  - Using local heap/stack-based source buffers has a performance impact
  - For small-message transfers – data size is critical to achieve peak performance
    – Understanding packet size and protocol selection is critical
  - Transitioning between protocol has performance impact
    – Solution - Use of SW protocol selection module
    – Need application specific inputs for better tuning
    – Highly visible with source buffers on local heap/stack

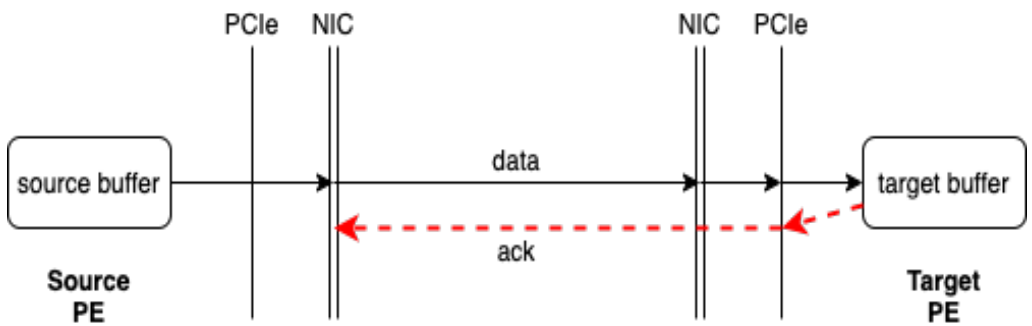
SLINGSHOT 11 NIC FEATURE – EVENT COMPLETION SEMANTICS

- Determines when an MPI RMA operation is completed
  - When can reuse source buffer
  - When the data is guaranteed to be visible on the target memory

- *Types of completions supported by Slingshot 11



Local Completion



Remote target completion

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

MPI PASSIVE RMA SUPPORTED COMPLETIONS

- MPI RMA (passive) supported completions

	Local completion	Remote target completion
MPI_Put	Return from **MPI_Win_flush_local** or **MPI_Win_flush_local_all**	Return from **MPI_Win_flush** or **MPI_Win_flush_all**

- Is it bad to use remote target completions?
  - No
  - Local completions are used to achieve better communication/computation overlap on pipelined algorithms
  - Scenarios where target visibility is required – remote target completions are still useful

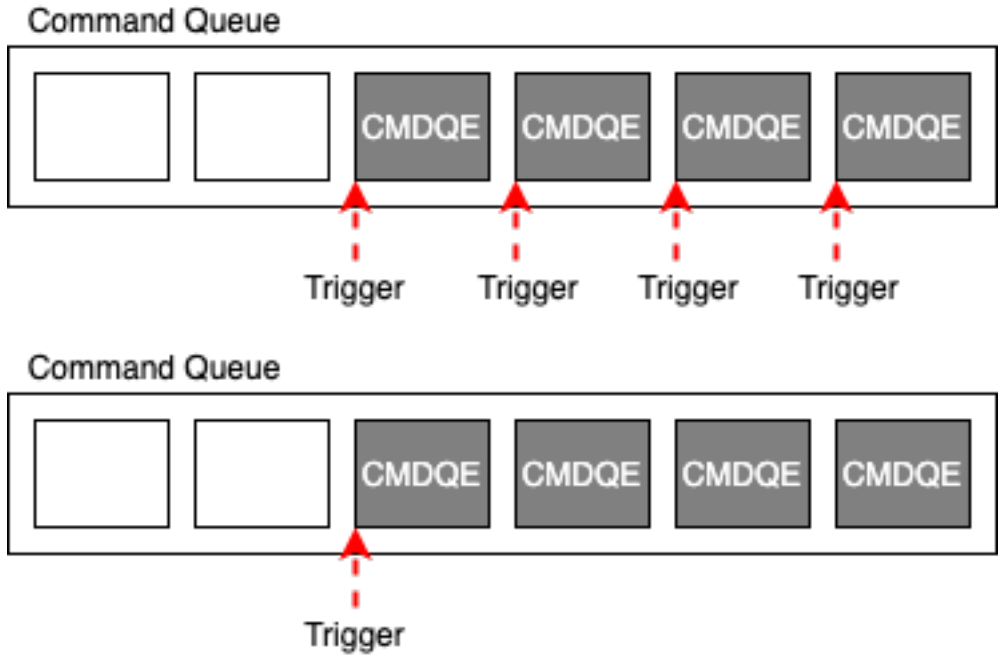
- There is a mismatch in the requirements
  - All completion semantics need to be specified during event launch
  - Not during event completion verification like flush/synch
  - Conservatively - implementations must implement everything as ”locally completed” events
  - **Is there a better option available to specify completions during event launch?**

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



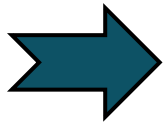
EXAMPLE FOR BUNDLING SUPPORT WITH OPENSHMEM

- No option to bundle OpenSHMEM operations
- 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);
```

- Need to adapt as epoch-based MPI RMA hints

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



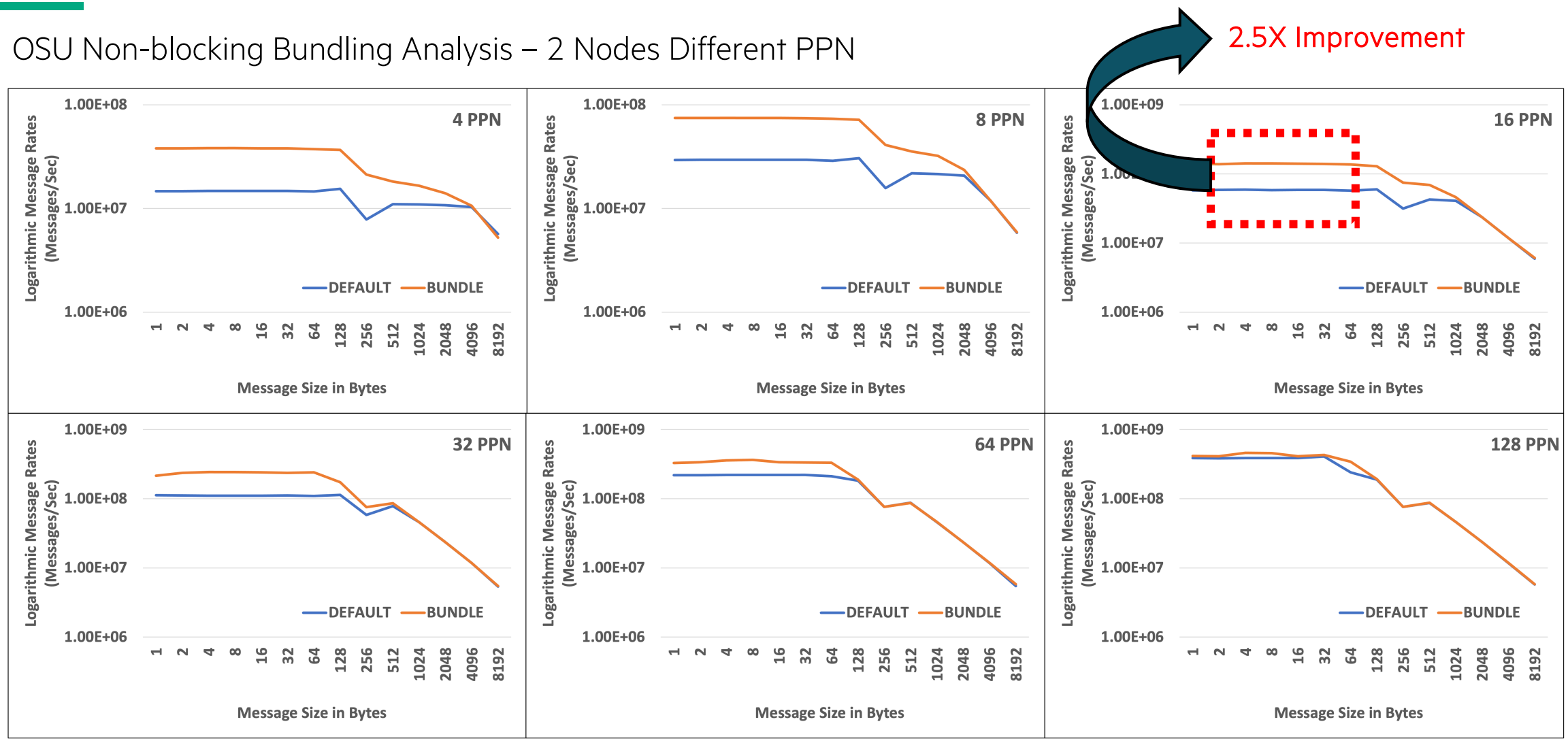
```
shmemx_ctx_session_start (SHMEM_CTX_DEFAULT,
SHMEM_SESSION_BUNDLE | SHMEM_SESSION_OP_PUT);

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



TRIGGERED COMMUNICATION OPERATIONS

- Deferred execution semantics
  - Event enqueued in the command queue list but deferred execution
  - Executed when certain conditions are satisfied

- Early thoughts
  - Our early experience is in using triggered operations for implementing put-with-signal in OpenSHMEM
  - Performance impact
    - Allowed communication/computation overlap
    - But – resource constrains – required to bundle PWS ops to use the same network resource

- How to use it in MPI RMA
  - Implicit implementation usage
    - Trigger operations during synchronization or completion points
  - Explicit user control
    - Not sure whether this is necessary
    - Is a persistent MPI RMA operation-like feature useful?

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

- Local completions can help to achieve better computation/communication overlap

- Bundling operations help use-cases with low PPN count
  - 8 PPN case shows 2.5X performance improvement
  - Drive better BW on low PPN case

- Triggered operations
  - Can be used implicitly
  - Need for explicit user interface is debatable

CONCLUSION

- Introduced and explored Slingshot 11 features that impacts one-sided communication operations

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

- What did we learn?
  - Slingshot 11 supports many knobs for performance tuning
  - Working to expose them in a user-accessible way
  - Adapting Slingshot 11 features for MPI RMA
    - MPI-RMA supports “most” of the features to expose Slingshot 11 specific features
    - Adapting these features is implementation design choices
    - Areas where we need updates includes – passing hints for bundling and completions

THANK YOU

nravi@hpe.com