Исследование структуры серии измерений задержки при нагрузочном тестировании коммутационной среды вычислительного кластера

Кафедра АСВК,
лаборатория Вычислительных комплексов
вед.науч.сотр. Сальников А.Н., студ. Волчанинов А.П.

Описание задачи

- Работа реализуется в рамках проекта clustbench (размещен на github)
- Набор средств для тестирования сети
- При тестировании происходит сбор данных о задержках (Ts_2-Ts_1) при передаче данных между MPI-процессами
- Доступны разные виды тестирования, в т.ч. one_to_one и all_to_all
- Информации о задержках записывается в трехмерную матрицу в виде статистических параметров, посчитанных по серии измерений.
- Пользователем задаются: количество измерений задержки, начальная, конечная длина сообщений, шаг изменения длины

Глобальная задача:

Предложить методы автоматического определения количества измерений задержки

Задача данной работы:

Исследовать структуру собираемых данных на предмет однородности и цикличности/периодичности

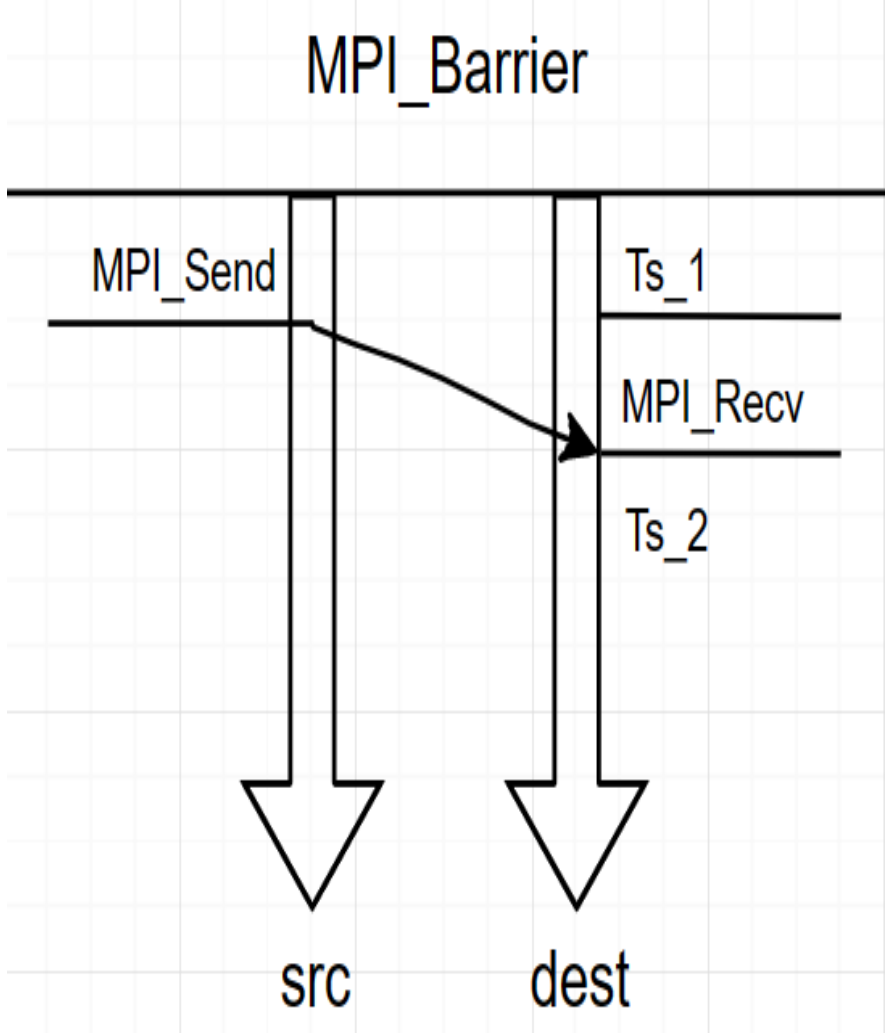


Схема измерения задержки в режиме one_to_one

Актуальность

Проект может быть интересен как администраторам суперкомпьютерных кластеров и инженерам подобных систем, так и программистам, работающим с распределенными вычислительными ресурсами. MPI-бэнчмарки позволяют:

- Получить представление о задержках и пропускной способности при передаче данных между различными узлами, процессорными ядрами и даже отдельными видеокартами, размещёнными на узле (узлах).
- Обнаружить проблемные линки.
- Верифицировать предполагаемую модель коммуникаций в сети, найти несоответствия и на основе всего этого прогнозировать накладные расходы на передачу данных по сети.
- Предоставить средство контроля за поведением наиболее важных функций MPI на конкретном сетевой архитектуре.

LinkTest

MVAPICH

Intel® MPI Benchmarks

MPI-бэнчмарки и статьи

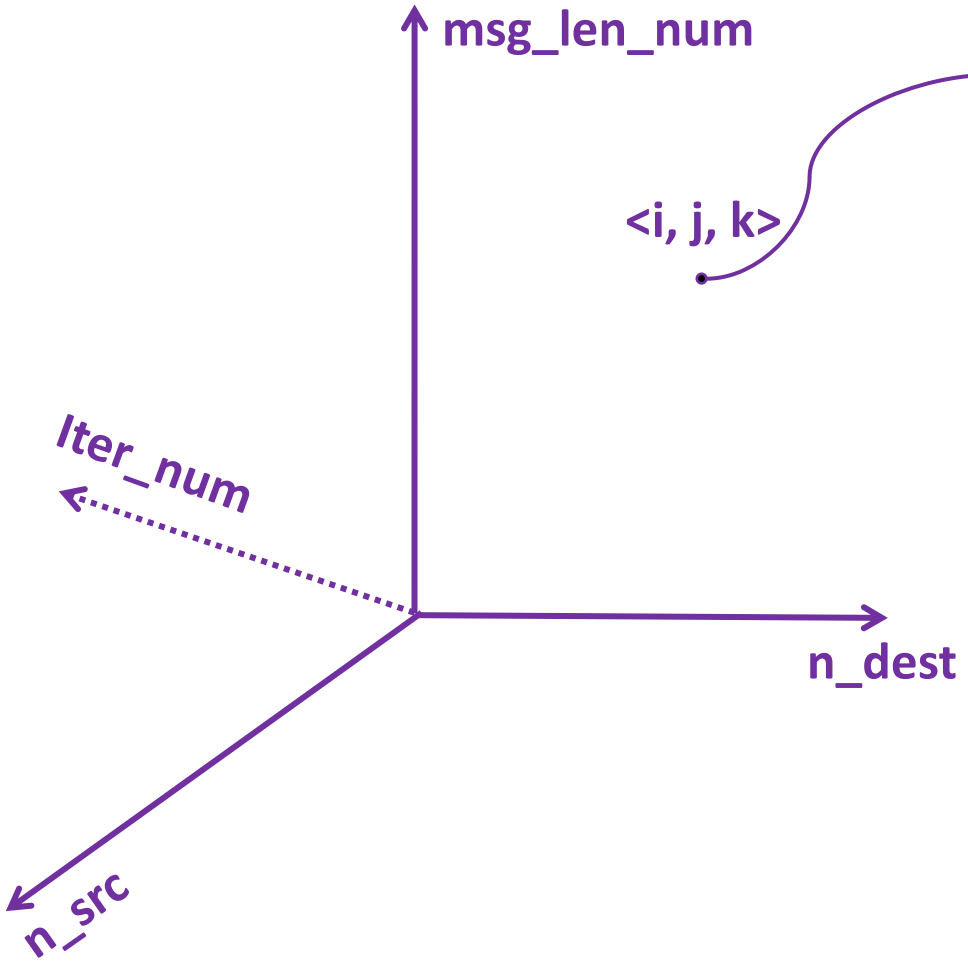
Можно выделить ряд MPI-бэнчмарков: LinkTest by JSC, MVAPICH by NBCL of The Ohio State University , Intel MPI Benchmark, и другие. Каждый из них ставит разные аспекты во главу угла, однако при изучении документации рассмотренных проектов, оказалось, что все, кроме Intel MPI Benchmark, оставляют вопрос определения количества итераций на откуп пользователю и предлагают ему самому определить статическое число итераций.

В статьях, написанных на основе проекта clustbench уже были проведены наблюдения за индивидулаьными значениями задержек. Установлен следующий факт: величины задержек хорошо описываются трёхпараметрич. логнормальным распределением. Однако, параметры этого распределения сильно отличаются при разных параметрах запуска теста.

MVAPICH

Intel® MPI Benchmarks

Формат сбора данных



Пусть пользователь задал:
Начальную длину len_beg, конечную длину len_fin
Число итераций = n, шаг изменения msg_len = s

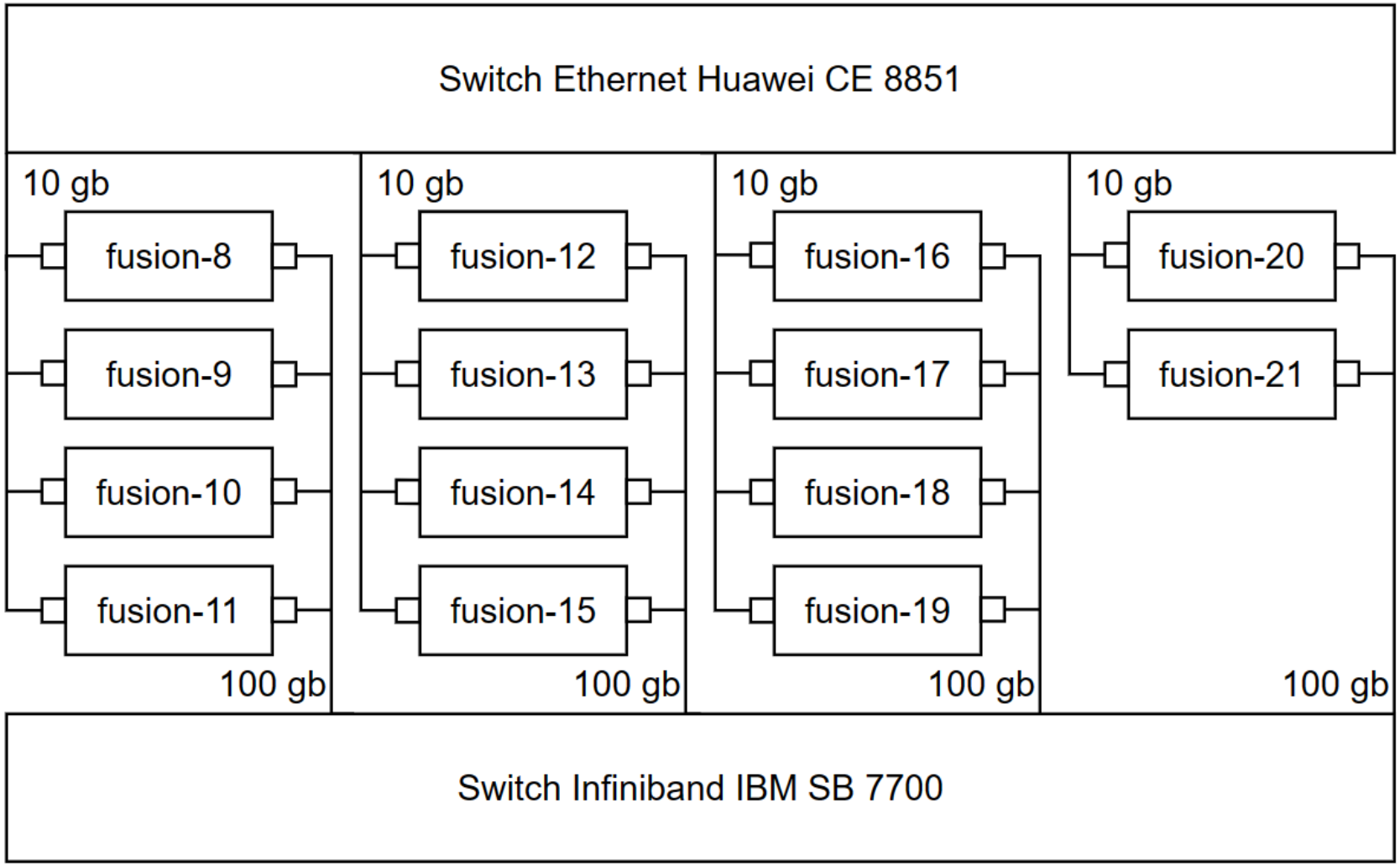
Процесс с номером i шлет процессу с номером j
n сообщений длины len_beg + k*s.

Для каждой посылки считается задержка и в итоге имеем вектор с n значениями задержки.
$[t_1, t_2, \ldots, tn]$.

По этим значениями вычисляем 4 параметра:
математическое ожидание (avg), дисперсию (dev), медиану (med), минимум (min).

В ячейку <i, j, k> записываем вектор <avg, dev, med, min>.

Стенд в ФИЦ ИУ РАН. InfiniBand.

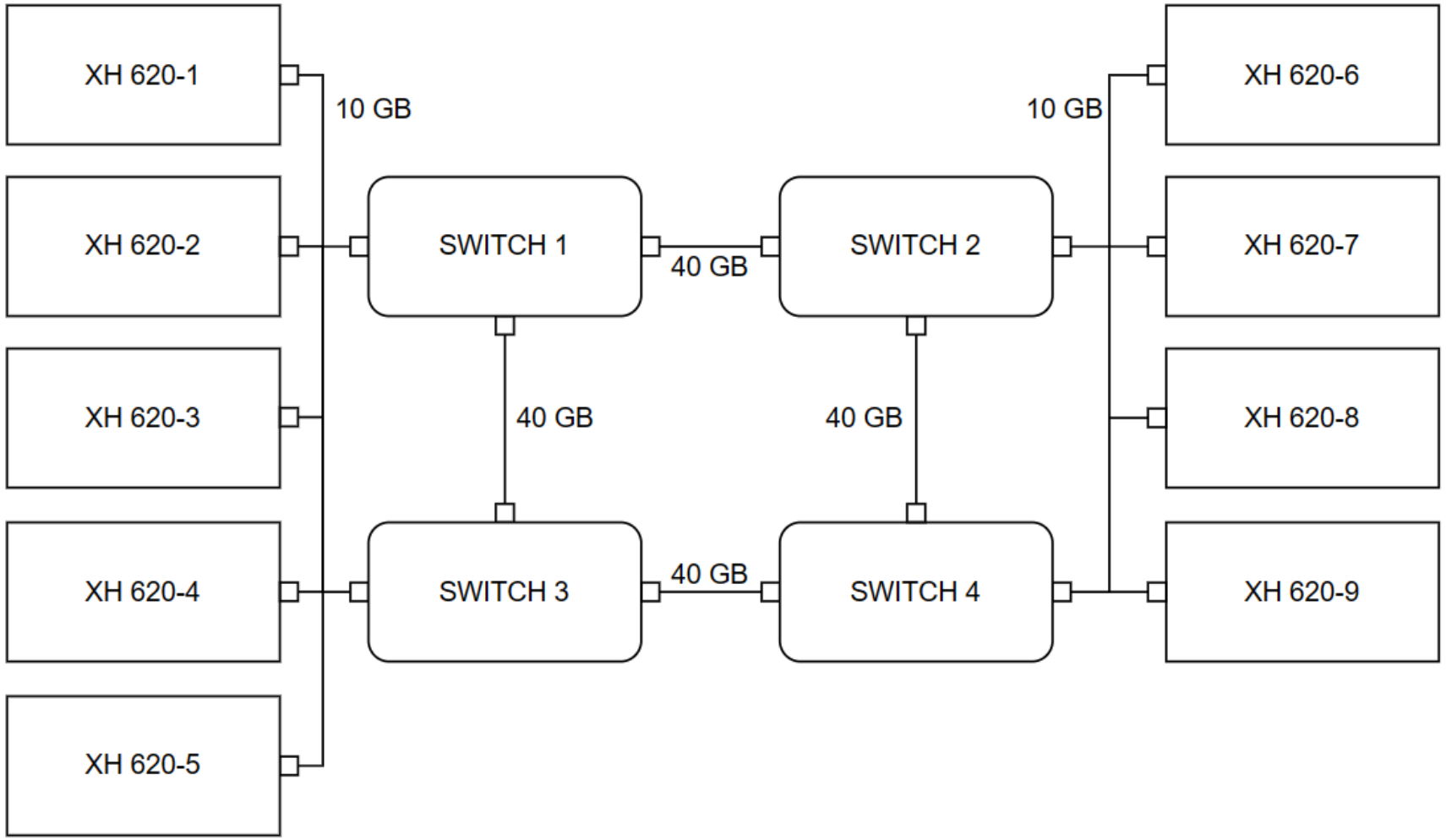


Высокопроизводительный вычислительный комплекс архитектуры Intel с IB EDR 100Gb/s (шина PCIe 4.0).

В состав комплекса входят 14 узлов на основе серверной платформы FusionServer 2288H со след. характеристиками:

2*CPU Intel Xeon Gold 6338 (2.0 GHz, 32 Core), 512 Gb RAM, 2*100G InfiniBand.

Стенд в ФИЦ ИУ РАН. Ethernet.

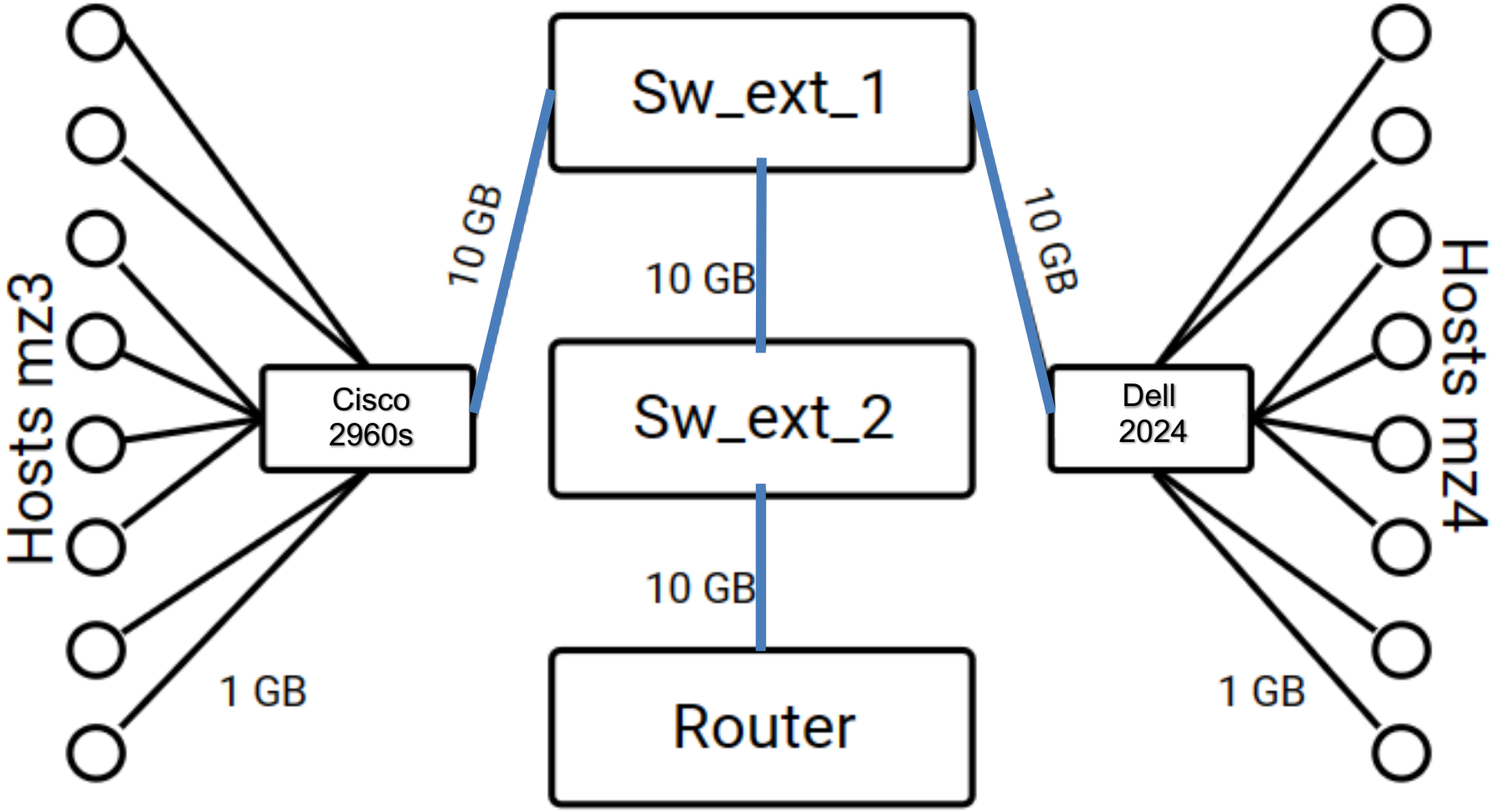


Высокопроизводительный вычислительный комплекс архитектуры Intel.

В состав комплекса входят 9 узлов на основе серверной платформы Huawei Server 23 XH620 со следующими характеристиками:

2*CPU Intel Xeon CPU E5-2683 v4 (2.1GHz, 16 Core), 512 Gb RAM, 2*10G Ethernet

Стенд в машинном зале ВМК МГУ



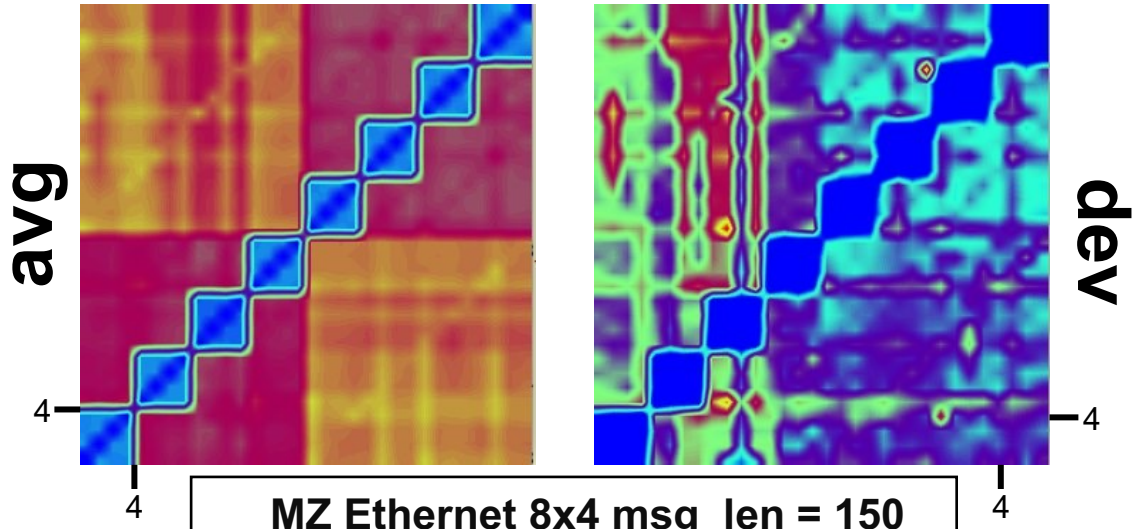
Модельный стенд в машинном зале.

В состав стенда вошли 16 машин со следующими характеристиками:
1*CPU Intel Xeon E3-1245 v5 (3.50GHz, 4 Core), 8 Gb RAM

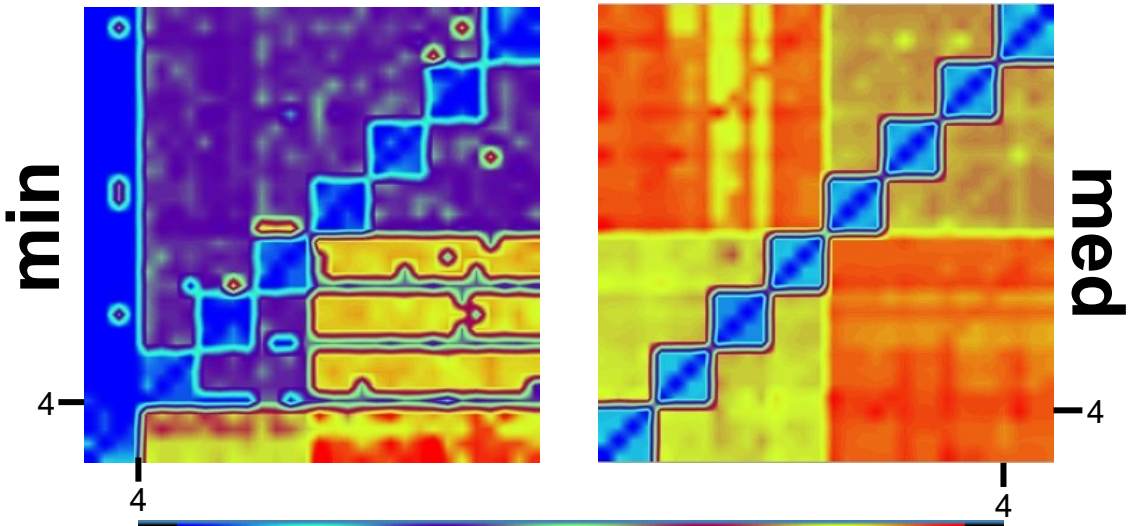
Оптика

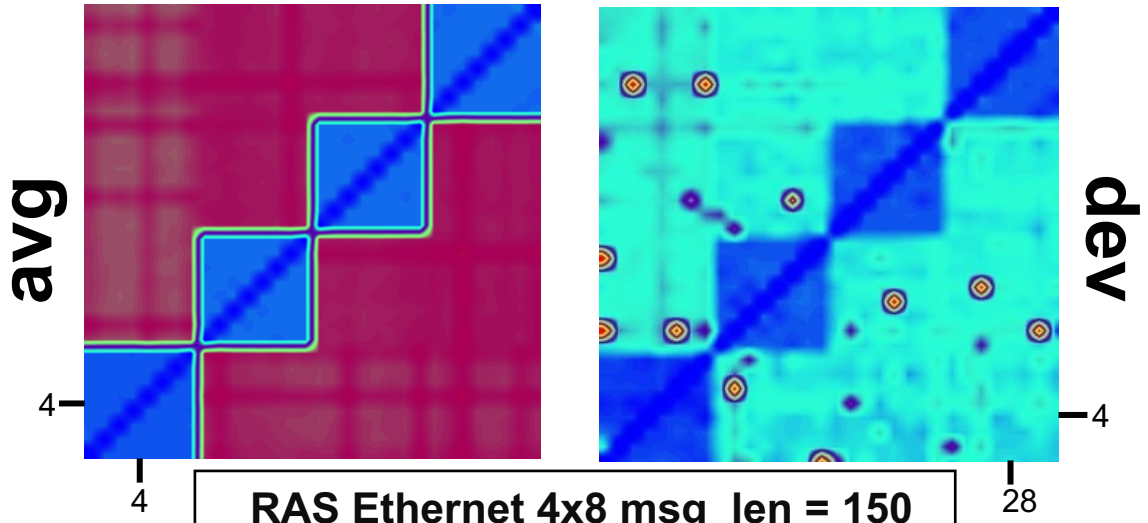
Патч-корд

Сравнение результатов для Ethernet

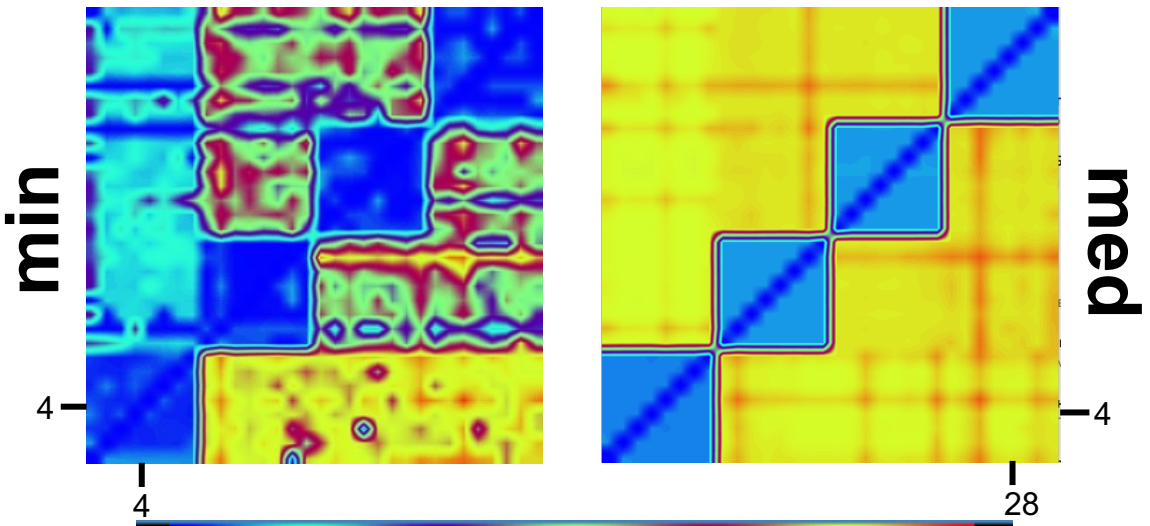


MZ Ethernet 8x4 msg_len = 150

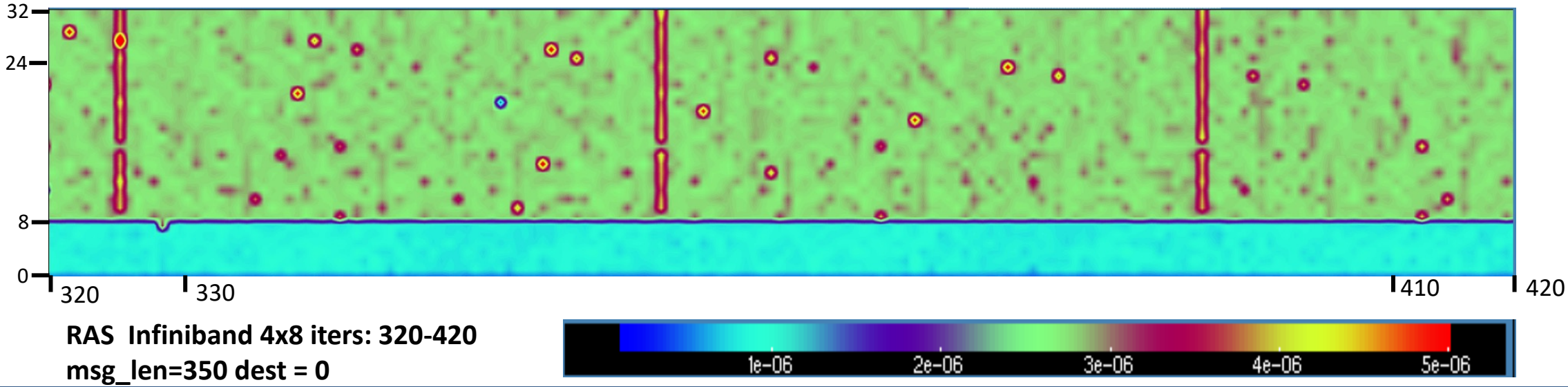




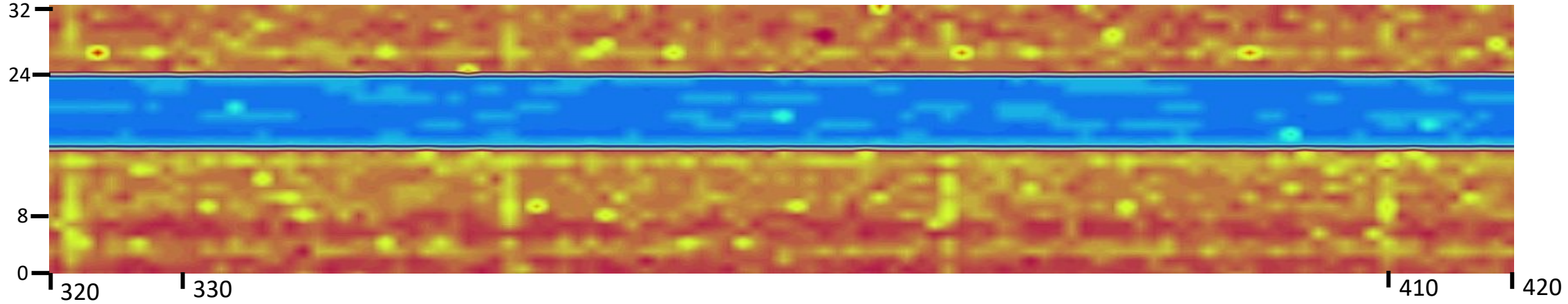
RAS Ethernet 4x8 msg_len = 150



Просмотр значений задержек (one_to_one)

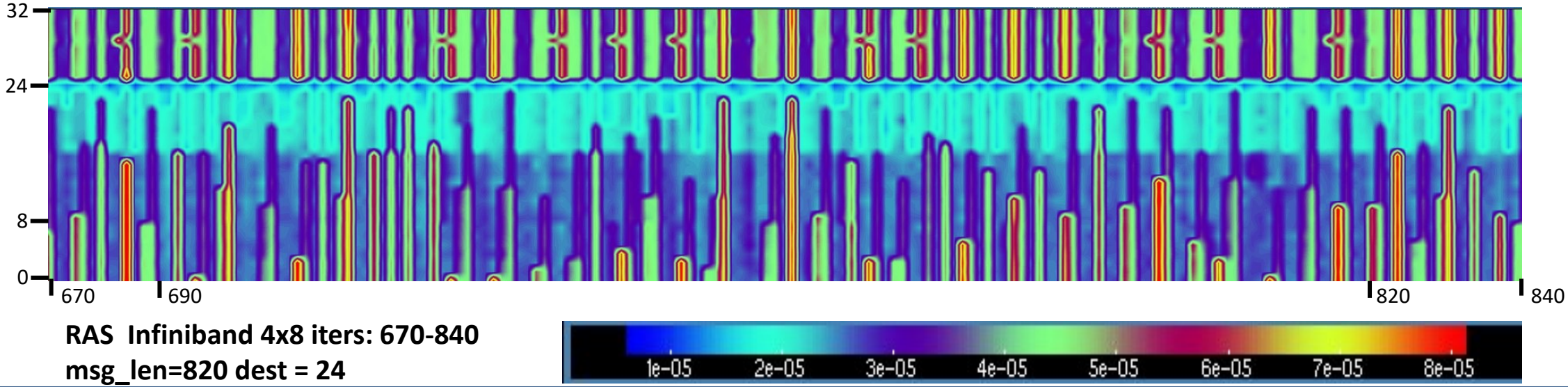


RAS Infiniband 4x8 iters: 320-420
msg_len=350 dest = 0

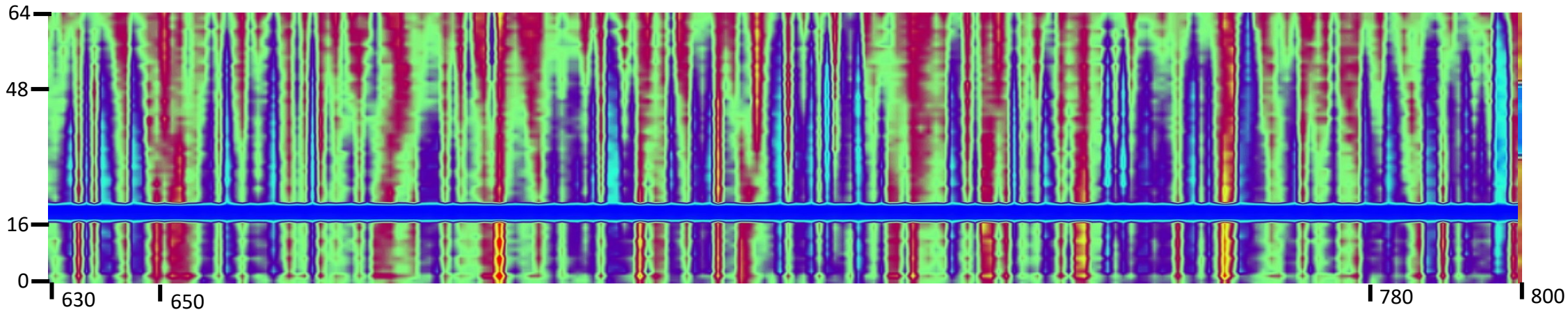


RAS Ethernet 4x8 iters: 320-420
msg_len=590 dest = 16

Просмотр значений задержек (all_to_all)

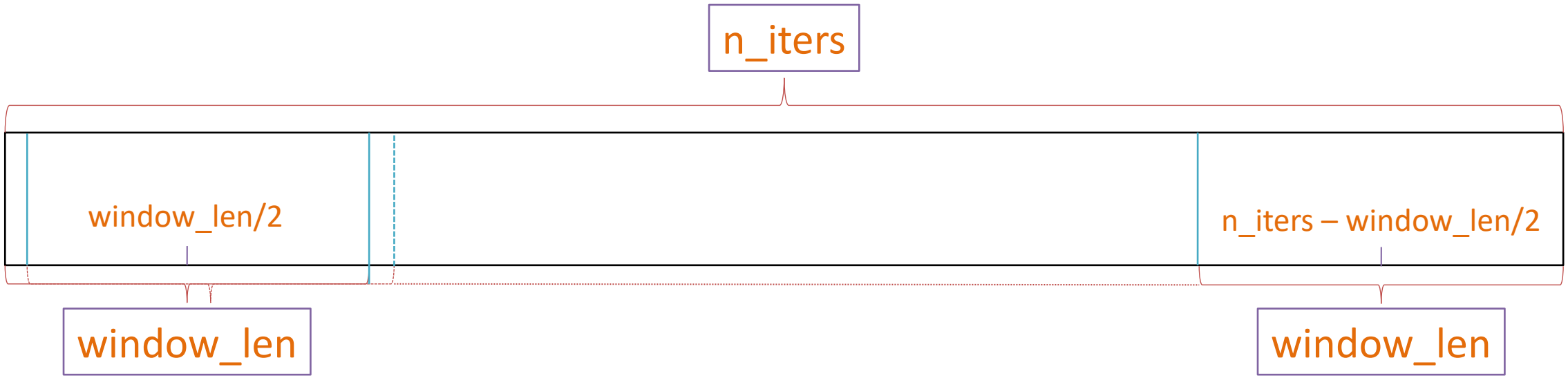


RAS Infiniband 4x8 iters: 670-840
msg_len=820 dest = 24



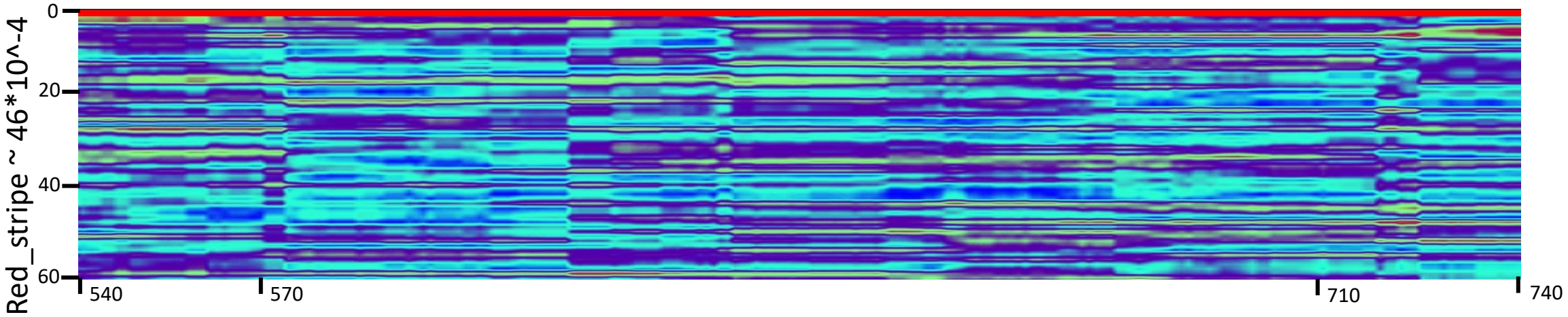
MZ Ethernet 16x4 iters: 630-800
msg_len=900 dest = 19

Применение дискретного преобразования Фурье

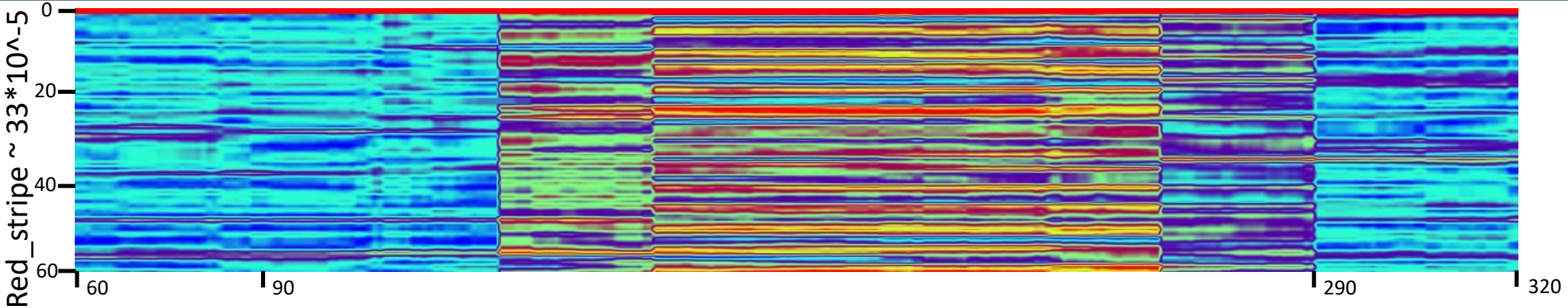


Итого: для каждой пары n_src, n_dest имеем (n_iters – window_len) окон

Просмотр мощностей гармоник (one_to_one)

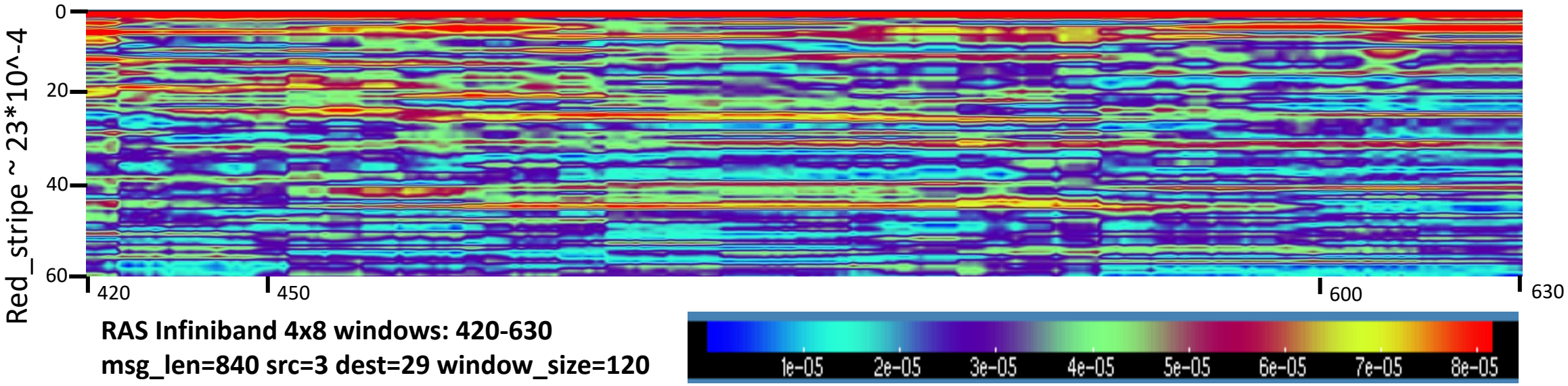


RAS Ethernet 4x8 windows: 540-740
msg_len=590 src=6 dest=24 window_size=120

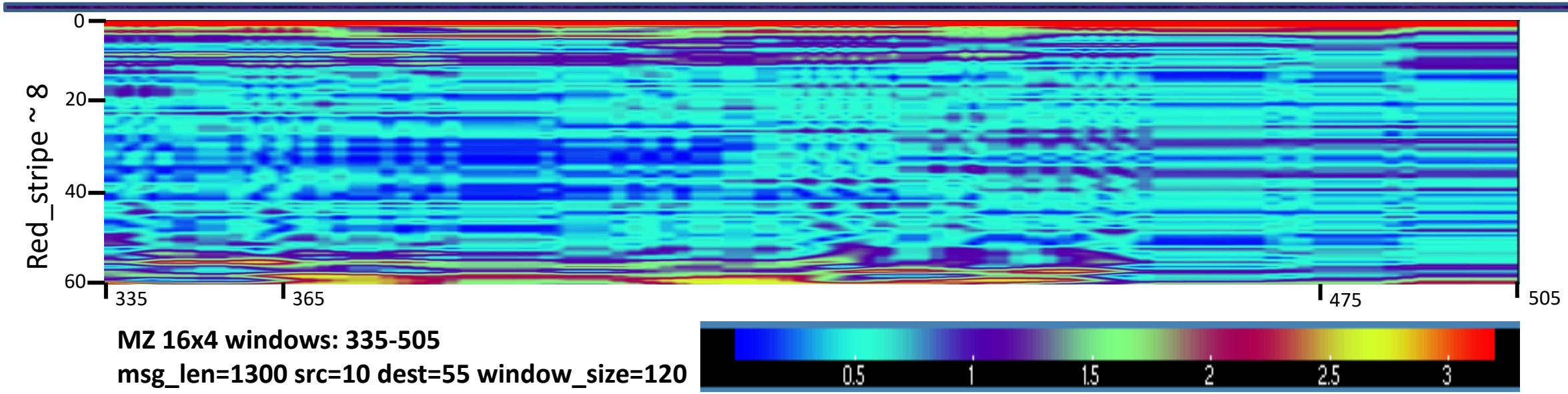


RAS Infiniband 4x8 windows: 60-320
msg_len=320 src=4 dest=28 window_size=120

Просмотр мощностей гармоник (all_to_all)



RAS Infiniband 4x8 windows: 420-630
msg_len=840 src=3 dest=29 window_size=120



MZ 16x4 windows: 335-505
msg_len=1300 src=10 dest=55 window_size=120

Результаты проделанной работы

- Собраны экспериментальные стенды для получения модельных данных:

  1) машинный зал ВМК МГУ. 2) комплексы ФИЦ ИУ РАН.

  Стенды позволили составить датасет, выложенный в систему Менделей (doi 10.17632/t4hn5d8wmx.2), который был использован в данной работе и будет служить основой для дальнейших исследований
- Модифицирован исходный код проекта clustbench, в результате чего появилась возможность сбора четырехмерной матрицы с индивидуальными значениями задержки с целью осуществления более подробного анализа наблюдаемых данных.
- Проведено подробное исследование структуры серии измерений задержек, показавшее цикличность и относительную стабильность данных.

Направления дальнейших исследований

- Разработка методов анализа периодических выбросов значений задержки.
- Поиск и реализация алгоритмов автоматизации сбора задержкив рамках проекта clustbench.
- Сравнение работы алгоритмов на разных сетевых инфраструктурах.
- Сравнение работы алгоритмов на разных режимах тестирования.
- Сравнение фактического времени исполнения тестирования.
- Совмещение данных алгоритмов с предварительным применением алгоритмов кластеризации графов.

СПАСИБО ЗА ВНИМАНИЕ!

Подробнее об Intel

- dynamic: Уменьшает количество итераций, когда задержка достигает максимального допустимого значения.
- multiple_np: Уменьшает количество итераций по мере роста размера сообщения.

all_to_all

1. При помощи MPI_Barrier происходит синхронизация между всеми процессами.
2. В каждом процессе ставится временная метка ts_{i0}, где i - номер процесса.
3. Все процессы используют MPI_Isend для посылки сообщение нужной длины всем процессам.
4. Все процессы n раз используют MPI_Irecv с опцией MPI_Waitany, где n - число процессов в тесте. Таким образом, будут получены сообщения от всех остальных процессов. После получения сообщения от процесса j в процессе i ставится временная метка ts_{ij} .
   Значение задержки между процессом j и процессом i вычисляется по формуле: ts_{ij}-ts_{i0}

Дискретное преобразование Фурье

Для поиска повторяющихся примерно через одинаковое число итераций событий написан модуль на python3.

На входе:

- Путь к файлу с четырёхмерной матрицей.
- Длина сообщения, которую необходимо зафиксировать.
- Длина окна для подсчета ДПФ.

Вычисления:

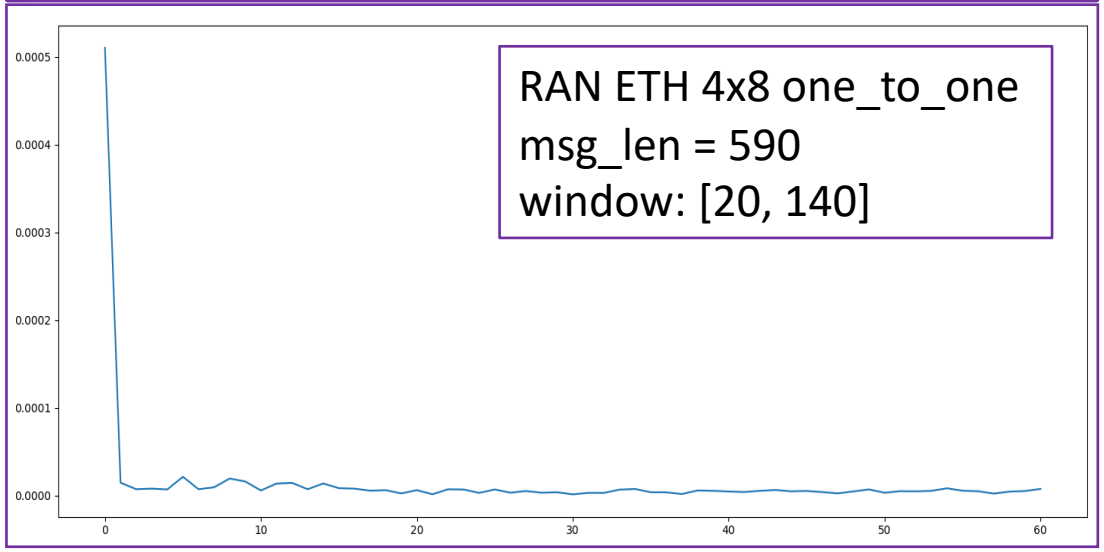
- Считаем мощности гармоник для всех окон длины N со смещением по вектору задержек.

На выходе:

- Файл в формате netcdf с четырехмерной матрицей с измерениями: <src, dest, pos, spek>.
- src – номер отправителя, dest – номер получателя, pos – номер окна, spek – номер гармоники.
- В ячейках матрицы хранятся измеренные мощности гармоник.

$$X_k = \sum_{n=-N/2}^{N/2-1} x_n e^{-\frac{2\pi i}{N}kn} k = -N/2, .., N/2-1$$

```
from scipy.fft import rfft, rfftfreq
…
yf = np.abs(rfft(window))
xf =rfftfreq(window_length, 1/window_length)
matplotlib.pyplot.plot(xf,yf)
```



Параметры запусков

- partition = intel (технология на которой основана сетевая инфраструктура: Ethernet) x 4 / intel_ib (технология на которой основана сетевая инфраструктура: Infiniband) x 4.
- name_of_test = one_to_one x 2 / all_to_all x 2 (для intel), для intel_ib – то же самое.
- nodes = 4/8 x 4. ntasks = 4/32 x 4. ntasks-per-node= 1/8 x 4.
- iter_num = 1000 x 8. beg_mes_len = 0 x 8
- end_mes_len = 1000 x 8. step = 10 x 8

Выводы

- В результате проведенных исследований во всех собранных данных наблюдались повторяющиеся паттерны. Более дорогие и совершенные сетевые инфраструктуры, ориентированные на работу в суперкомпьютерных кластерах, показали более однородные и цикличные результаты.
- При тестировании в режиме all_to_all наблюдаются частые скачки задержки. Выдвинуты две гипотезы о причинах подобного поведения.
- Предложенная модификация с возможностью сбора четырехмерной матрицы позволила отслеживать моменты переполнения буферов в сети на пути следования посылок — при достижении критического размера сообщения значения задержки равномерноувеличивались для всех пар задействованных процессов.
- Анализ мощностей гармоник позволил утвердить факт периодичности наблюдаемых данных.
- Основываясь на проведенных исследованиях было предложено два концептуально разных метода автоматического определения числа измерений задержки. При анализе алгоритма, основанного на оценке расстояний между векторами спектра, выяснилось, что факт наличия очень мощной гармоники, ограничивает объективность работы данного алгоритмы, в связи с чем была предложена его модификация. На всех данных, собранных с испытательных стендов были найдены параметры, при которых алгоритмы вели себя стабильно.
- В случае реализации данных алгоритмов в рамках проекта clustbench, пользователю будет дана потенциальная воз-можность: во-первых, уменьшить число итераций. Однако, чтобы воспользоваться данными возможностями, следует грамотно подойти к подбору параметров длястратегии определения выбора окон и точности.

Исследование структуры данных

- В процессе работы, система сбора информации была модифицирована: добавлена опция сохранения данных о всех собираемых измерениях в виде четырёхмерной матрицы.
- Написан модуль для получения среза в 4D матрице по указанной длине сообщения.
- Настроен экспериментальный стенд в машинном зале.
- Собраны данные с комплекса ФИЦ ИУ РАН и стенда в машинном зале.