УДК 004.272.43+004.23+004.25

**Высокоуровневая система параллельного программирования сети многоядерных гибридных процессоров**

Биконов Дмитрий Владиленович,

Черников Александр Владимирович,

Павлов Павел Алексеевич,

Дадашев [Магомедвели Сийидгусенович](mailto:m.dadashev@module.ru),

Бражкин Алексеевич Алексеевич,

Пузиков Артем Дмитриевич,

Сивцов Алексей Сергеевич,

Зимин Петр Алексеевич,

Эйсымонт Леонид Константинович

*ЗАО НТЦ "Модуль".125190, а/я 166. г. Москва, 4-ая улица 8 Марта, д.3. d.bikonov@module.ru*

**Аннотация**

Рассматриваются многопроцессорные платы на базе процессоров-ускорителей GPU Ampere, Gaudi и Colossus для узлов суперкомпьютеров, варианты многопроцессорных плат с процессорами-ускорителями NeuroMatrix NM6408MP (1879ВМ8Я), вопросы разработки для таких плат библиотеки управления процессами и их взаимодействиями БУПВ/МХ на базе ранее разработанной библиотеки БУПВ/М для одного NM6408MP.

**Ключевые слова:** многопроцессорные узлы суперкомпьютеров; процессор-ускоритель; межпроцессорные коммуникации; коммутация каналов; коммутация пакетов.

# **Введение**

Процессор NeuroMatrix NM6408MP(1879ВМ8Я) [1] включает 21 узел, из которых 5 узлов с суперскалярными ядрами на базе Cortex A5 и 16 узлов со скалярно-векторными ядрами собственной разработки (ядра NMC4). Узлы содержат локальные памяти с расслоением 4 и 8, соответственно с объемом 256 KB (суперскалярные) и 512 КB (скалярно-векторные и центральный суперскалярный). Пиковая производительность процессора, определяемая скалярно-векторными узлами, работающими на частоте 1 GHz, составляет 0.125 TF (FP64, 64-разрядные числа с плавающей точкой) и 0.5 TF (FP32, 32-х разрядные числа с плавающей точкой)). Энергопотребление не более 20 Вт. Процессорные узлы разбиты на четыре кластера, один кластер – это один суперскалярный узел и четыре скалярно-векторных. Имеется один центральный суперскалярный узел. Все суперскалярные узлы имеют интерфейс с DDR3 памятью с пропускной способностью каждого 6.4 GB/s, общий объем подключаемой памяти – до 5 GB. Для межпроцессорных коммуникаций имеются 4 дуплексных линка (EL-линки) с односторонней пропускной способностью каждого линка до 2 GB/s. Для связи с хост-процессором включен PCIe 2.0 интерфейс с дуплексной пропускной способностью 4 GB/s. Кроме этого реализованы интерфейсы Ethernеt 10/100 и SPI и другие [1].

Структура платы с одним NM6408MP представлена на рисунке (см. Рисунок 1(а)) [2]. Такая плата вставляется в один разъем PCIe материнской платы персонального компьютера (далее – ПК) с хост процессором. Для ПК имеется возможность подключения до пяти таких плат с объединением их соединительной кросс-платой по топологии “каждый с каждым” Можно считать все это 5-процессорной платой для ПК, подключаемой к материнской плате через 5 разъемов PCIe. Схема такого подключения процессоров NM6408MP к ПК показана на рисунке (см. Рисунок 1(б)).

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\final\рис1.tif

Рисунок 1. Схема платы NM6408MР и схема подключения пяти плат NM6408MP к персональному компьютеру

Для однопроцессорной платы разработаны низкоуровневые библиотеки загрузки и обмена данными с хост-процессором, низкоуровневые библиотеки функций синхронизации и коммуникации процессов низкого уровня БФСКН-Ф и БФСКН [3], а также высокоуровневая библиотека управления процессами и их взаимодействиями БУПВ/М [3, 4], которая поддерживает модели параллельных программ типа MPI и PGAS, а также графовые модели вычислений с управлением потоком данных.

Разработка процессоров-ускорителей, многопроцессорных плат и суперкомпьютерных узлов на их основе – сейчас главное направление в области суперкомпьютерных технологий, что особенно актуально, для систем искусственного интеллекта. Начиная работы по этому направлению, естественно оценить свои возможности и достижимые в краткосрочной и дальней перспективе цели, сопоставив их с мировым уровнем. Рассмотрению этого посвящена первая часть данной статьи.

Для пятипроцессорной платы в настоящее время применимы все библиотеки нижнего уровня, а расширенный вариант библиотеки БУПВ/М для такой платы и более мощных плат для суперкомпьютерных серверных узлов (далее – БУПВ/МХ) предстоит разработать, этой проблеме и посвящена данная статья в ее второй части.

Имеющийся в настоящее время вариант 5-процессорной платы с NM6408MP обладает пиковой производительностью 1 TF (FP64) и 2.5 TF (FP32), а общий объем внекристальной памяти на этой плате составляет 25 GB. Если сравнивать с зарубежными образцами, которые рассмотрены далее, то это очень мало. Но если обратить внимание на баланс пиковых производительностей хост-процессоров и многопроцессорных ускорительных плат, которые к ним подключаются, то оказывается не все так плохо. Применение плат с NM6408MP (причем для серверных узлов) предполагается в сочетании с отечественными универсальными процессорами, в настоящее время это, например, Эльбрус 8СВ c пиковой производительностью 0.25 TF (FP64) и 0.5 TF (FP32). Таким образом, подключаемая к ПК с таким процессором плата добавит, например, 5-кратную производительность на FP32. Возникает вопрос, насколько это хорошо, это далее рассмотрено в разделе с использованием результатов анализа зарубежных плат.

Далее в разделе 2 рассмотрены варианты зарубежных плат с процессорами-ускорителями и варианты многопроцессорных плат с NM6408MP. После этого в разделе 3 рассмотрены задачи, которые предстоит решить в первую очередь при создании БУПВ/МХ на базе БУПВ/М.

# **Многопроцессорные узлы с процессорами-ускорителями**

В данном разделе приведены основные особенности многопроцессорных плат с процессорами ускорителями GPU Volta [5,6] (DGX-1, DGX-2), GPU Ampere [7] (DGX), нейропроцессорами Gaudi [8,9] (HLS-1), нейропроцессорами Colossus Mk2 GC200 [10, 11] (M2000). Эти платы входят в состав суперкомпьютерных узлов.

Платы с процессорами Gaudi и Сolossus имеют значительную перспективу, поскольку за счет использования линков процессоров и линков 100G Ethernet позволяют строить суперкомпьютеры с соединениями “host-less” типа, без использования подключаемых к хост-процессорам дорогостоящих и энергоемких межузловых сетей типа Mellanox Infiniband или других такого класса.

# **Суперкомпьютерные узлы на базе GPU Volta и GPU Ampere**

Появление GPU Volta и Ampere предоставило новые возможности разработки мощных вычислительных узлов на их основе для построения суперкомпьютеров транспетафлопсного класса, приближающихся вплотную к экзафлопсному уровню производительности. В системах ИИ в последние годы оказалась крайне востребованной потребность именно в суперкомпьютерах такого уровня. В подготовленной авторами таблице 1 приведены характеристики нового вычислительного узла DGX A100 [7] в сравнении с ранее разработанными на базе GPU Volta узлами DGX-1 [5] и DGX-2 [6]. Схемы соединения компонентов этих узлов представлены на рисунках (см. Рисунок 2 и Рисунок *3*).

Следует обратить внимание на следующее:

* внутриузловые сети основаны на применении высокоскоростных NVLink-ов (дуплексная пропускная способность каждого из них достигает 50 GB/s), а также многопортовых коммутаторов NVSwitch, множества коммутаторов PCIe 3.0/3.1 и 4.0 с пропускной способностью до 35 GB/s;
* внекристальная внутриузловая память имеет три уровня: первый – HBM-память среднего объема в несколько сотен гигабайт, но с высокой пропускной способностью; второй – DDR4-память объемом до терабайтов на платах хост-процессоров; третий – очень большая память в десятки терабайт на уровне хост-процессора на SSD-дисках;
* интерфейс PCIe с хост-процессорами имеется у каждого процессора-ускорителя;
* межузловые сети строятся на базе маршрутизаторов фирмы Mellanox с пропускной способностью линка до 200Gb/s и высокоскоростного GbE с пропускной способностью 100-200 Gb/s, причем топология межузловой сети имеет тип Dragonflу, с ограниченным диаметром.

Количество процессоров-ускорителей в узле 8 или 16. Рассмотрение топологии их соединений показывает, что если в первой плате с DGX-1 (см Рисунок 2 (а)) применялось прямое соединение посредством мощных NVLink-ов, то потом от этого отказались. Стали использоваться коммутаторы NVSwitch (см. Рисунок 2 (б)), причем если в DGX-2 16 GPU Volta были подключены к 12 8-портовым коммутаторам (96 портов, по 6 портов на GPU), то в следующей разработке DGX 8 GPU Ampere подключены к 8 8-портовым коммутаторам (64 порта, по 8 портов на GPU). Видно, что роль сетевых соединений процессоров ускорителей возрастает.

Вместе с тем, межузловая сеть доступна только через хост-процессор, это приводит к лишним накладным расходам, что, как далее будет показано, негативно влияет на масштабируемость производительности при увеличении количества узлов, используемых при решении задач.

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\final\рис2.tif

Рисунок 2. Схема соединения узлов DGX-1 и DGX-2

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\рис4.tif

Рисунок 3. Схема соединений узла DGX A100

Уже известно об использовании фирмой NVIDIA узлов DGX A100 в построенном суперкомпьютере с пиковой производительностью до 700 PF, а также другие примеры построения суперкомпьютеров такого уровня на этих узлах.

# **Суперкомпьютерные узлы на базе нейропроцессоров Gaudi и Сolossus**

Вычислительные узлы с многопроцессорными платами c процессорами-ускорителями в виде GPU имеют широкий спектр применений. Однако в такой важной в настоящее время области, как нейровычисления, таким платам серьезнейшую конкуренцию стали составлять платы со специализированными на нейровычисления процессорами-ускорителями. Одни из наиболее удачных – это Gaudi [8, 9] для задач обучения нейросетей и Сolossus [10, 11] как для обучения, так и применения (задачи “вывода”) нейросетей.

Данные по таким конкурирующим с GPU более простым специализированным под нейровычисления процессорам, особенно по многопроцессорным платам на их основе, в данной статье интересны из-за того, что NM6408MP является процессором-ускорителем такого же типа. Публикуемые по Gaudi и Сolossus результаты сравнительного с GPU Volta оценочного тестирования впечатляют [8, 13].

В работе [8] приводятся результаты сравнения Gaudi с GPU Volta, часть их приведена на рисунке (см. Рисунок 4) для теста обучения нейросети ResNet-50 (обработка изображений). Один процессор Gaudi оказался на 14% более производительным (см. Рисунок 4(а), ResNet-50 IPS) и более, чем в 2 раза энергоэффективным (см. Рисунок 4(а), ResNet-50 IPS/W). В контексте данной статьи более интересны результаты по масштабированию производительности. На рисунке (см. Рисунок 4(б)) приведены данные по деградации производительности одного процессора при их увеличении до 512. Видно, что для сети на процессорах Gaudi результаты однозначно лучше. Использовались 4-х процессорные платы (см. Рисунок 4(б)).

Результаты сравнительного с GPU Volta оценочного тестирования первого поколения процессоров Colossus на 2-х процесcорных платах С2 имеются в [13]. Они показывают преимущество процессоров этой платы особенно на задачах обучения в разы.

Результаты прямого сравнения на задачах Colossus Mk2 GC200 c GPU Ampere пока не найдены, но по пиковым характеристикам сравнивают один узел DGX (см. Рисунок 3) с восемью 1U узлами M2000 (cм. Рисунок 6). Восемь узлов M2000 по производительности на FP32 превосходят DGX в 12 раз, по AI-производительности на FP16/FP32 превосходят в 3 раза, по AI-памяти (HBM- или DDR-паять на платах) превосходят в 10 раз. При этом стоимость DGX — $199 000, а восьми плат M2000 — $256 000. Реальная производительность Colossus второго поколения превосходит Colossus первого поколения на задачах обучения и вывода более, чем в 7–9 раз.

Сравнительные характеристики плат приведены в Таблице 1.

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\gaudi-volta.tif

Рисунок 4. Сравнение процессоров-ускорителей Gaudi и GPU Volta (V100)

**Таблица 1. Сравнительные характеристики многопроцессорных плат с процессорами-ускорителями, баланс характеристик**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Характеристика** | **Процессор ускоритель** | | | | | |
| **V100**  **Volta** | **V100 Volta** | **A100**  **Ampere** | **Gaudi**  **HL-2000** | **Colossus (2018)** | **Colossus**  **GC200** |
| AI — производительность процессора (FP16/FP32) | 120 TF | 120 TF | 312 (624) TF | 130 TF | 124.5 TF | 250 TF |
| Количество линков  процессора | 6 NVLink | 6 NVLink | 12 NVLink | 10 линков  100GbE | 10 IPU-линков | 10 IPU-линков |
| Кол-во процессоров-ускорителей на плате | 8 | 16 | 8 | 1 (HL-205) | 2 (С2) | 4 (ML-2000) |
| Кол-во плат с процессорами-ускорителями в ВУ | 1 (DGX-1) | 1 (DGX-2) | 1 (DGX) | 8 (HLS-1) | 8 (DELL DSS 8440) | 8 |
| Общее кол-во процессоров-ускорителей в ВУ | 8 | 16 | 8 | 8 | 16 | 64 |
| Σ PACC | 960 TF | 1920 TF | 2496 (4932) TF | 1040 TF | 1992 TF | 8000 TF |
| Хост-процессоры ВУ, общее количество ядер и тредов, Σ PHOST — общая пиковая производительность хост-процессоров (FP32) | 2хXeon  Platinum 8168,  40 ядер  80 тред  8 TF | 2хXeon  Platinum 8168,  48 ядер  96 тред  9.6 TF | 2xAMD Rome  7742,  128 ядер  256 тред  14 TF | н/д | 2хXeon  Platinum 8168,  48 ядер  96 тред  9.6 TF | н/д |
| Баланс Σ PACC / Σ PHOST | ~ 120 | ~ 200 | ~ 180(360) | ~100  (оценка) | ~ 200 | ~ 800  (оценка) |
| Σ VСASH/SMEM | 0.256 GB  (CASH) | 0.512 GB  (CASH) | 0.648 GB  (CASH) | н/д | 4.864 GB  (SMEM) | 57.6 GB  (SMEM) |
| Σ VHBM/DDR4 — суммарная внекристальная память | 256GB  (HBM) | 512GB  (HBM) | 320GB  (HBM) | 256GB  (HBM) | нет | 3840GB  (DDR4) |
| Баланс  Σ VHBM/DDR4 /Σ VСASH/SMEM | 1000 | 1000 | 490 | н/д | нет | ~ 67 |
| Баланс  Σ PACC / Σ VHBM/DDR4 | 3.7 TF/GB | 3.7 TF/GB | 7.8(15.6) TF/GB | 4 TF/GB | нет | 2 TF/GB |
| DDR-память  хост-процессоров | 51GB | 1000GB | до 2000GB | н/д | 768GB | н/д |
| SSD-память  Хост-процессоров | 7.7TB | 30-60TB | 15-30TB | н/д | 16TB | н/д |
| Коммуникационная сеть между вычислительными узлами | 4хInifiniband  Mellanox EDR + 2x10 GbE | 8хInifiniband  Mellanox EDR + 8x100 GbE | 8хInifiniband  Mellanox HDR + 2x100 GbE | 24х100 GbE  (host less) | ? x 10/25  GbE  (host less) | 4 IPU-link + 2x100 GbE  (host less) |
| Энергопотребление вычислительного узла | 3.5 KW | 10 KW | 6.5 KW | 3 KW | 4.8 KW | н/д |

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\рис6.tif

Рисунок 5. Топология соединений и внешний вид серверной платы с нейропроцессорами Gaudi

Принципиальная особенность процессоров Colossus первого и второго поколения — огромное количество 6-тредовых тайлов (соответственно 1216 и 1472), каждый из которых имеет локальную статическую память с высокой пропускной способностью за счет сильного расслоения и средней мощности блок ускорения тензорных операций с матрицами. Наличие огромной внутрикристальной памяти значительно отличает эти процессоры от остальных, см. строку “**ΣVСASH/SMEM** — суммарная внутрикристальная память процессоров-ускорителей плат вычислительного узла” в таблице 1. Топология и внешний вид 4-х процессорной платы с Colossus второго поколения даны на рисунке (см. Рисунок 6).

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\рис7.tif

Рисунок 6. Топология соединений и внешний вид серверной платы с нейропроцессорами второго поколения Colossus GC200

Интересен и вариант реализации платы С2 с процессором-ускорителем первого поколения, ее внешний вид и вид серверного узла с восемью платами С2 представлен на рисунке (см. Рисунок 7).

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\final\рис7.tif

Рисунок 7. Внешний вид 2-х процессорной платы С2 и суперкомпьютерного узла с 16 процессорами Colossus первого поколения

По приведенным общим характеристикам Gaudi и Colossus процессор NM6408MP имеет некоторые общие с ними архитектурные черты, а исследуемый в настоящее время вариант его развития, процессор mPX [12], - тем более. Однако сильное преимущество этих процессоров, которым не обладают большинство других, — это наличие в большом количестве линков межкристльных взаимодействий, см. строку “Количество линков процессора” в таблице 1. Наиболее мощные линки у Colossus — дуплексные, с пропускной способностью 64 GB/s. Процессор NM6408MP имеет только 4 межпроцессорных линка (EL-линки), дуплексная пропускная способность каждого из них по 4 GB/s. Ясно, что это надо улучшать, особенно в проекте по mPX[12].

Большое количество межкристальных линков позволяет реализовать сильные связи на многопроцессорных платах (см. топологии плат на Рисунок 5 и Рисунок *6*), обойтись без межузловых сетей, подключаемых к хост-процессорам.

Систематизированные в таблице 1 характеристики по платам позволяют оценить и сравнить их сбалансированность, это имеется в строках со словом “Баланс” в наименовании. Это своеобразные показатели качества, такие оценки сбалансированности давно применяются в вычислительной технике. На основе одной из таких оценок и был сделан вывод о целесообразности разработки многопроцессорных плат с процессорами-ускорителями NM6408MP для серверных узлов с хост-процессорами в виде отечественных универсальных микропроцессоров уровня производительности Эльбрус 8CB. Рассмотрим такие оценки производительности, их три.

1. *Показатель сбалансированности* **ΣPACC/Σ PHOST**, отношение суммарной производительности процессоров-ускорителей к производительности хост-процессоров исчисляется сотнями и увеличивается в новых разработках. Для 5-ти процессорной платы с NM6408MP такой показатель равен 5, для связки из 8-ми 8-процессорных серверных плат (похоже на решение с платами M2000 c процессорами Colossus) такая оценка равняется 32-м, а для связки из 8-ми 16-процессорных серверных плат этот показатель имеет значение 64. Собственно говоря, это и объясняет, в основном, целесообразность разработки серверных плат c NM6408MP.

Далее перейдем к оценкам сбалансированности, включающим объемы внекристальной и внутрикристальной памяти. Отметим, что подключенная к одному NM6408MP DDR3-память объемом 5 GB вызывала всегда некоторую тревогу, не мало ли? Исторически сложилось, что производительность процессора связывается с объемом памяти, доступной к нему. По этой причине рассмотрим следующую оценку сбалансированности.

1. *Показатель сбалансированности* **ΣPACC/ΣVHBM/DDR4.** Большие значения этого показателя должны вызывать опасения, что данных для процессора в памяти будет не хватать. Для четырех из пяти плат этот показатель 2-4 TF/GB. Только на плате с GPU Ampere он вырос до 7.8 (значение 15.6 – расчетное). Отметим, что такая внекристальная память необходима. Это показывает опыт фирмы Graphcore c ее процессорами Colossus первого и второго поколения. На серверной плате с 8-ю 2-х процессорными платами С2 такой памяти нет, но на более поздних платах М2000 такая память была введена, причем большого объема — на 4 процессора приходилось 480 GB DDR4-памяти. Для этого даже была разработана специальная микросхема — мост IPU Gateaway. Для этих плат оценка такой сбалансированности стала 2 TF/GB.

Что касается плат с NM6408MP, то такой показатель получается равным 0.5 TF/5 GB = 0.1 TF/GB. Это говорит о том, что внекристальной памяти для такой пиковой производительности процессоров вполне достаточно, опасения были напрасными.

Обращения к внекристальной памяти имеют времена выполнения до сотни тактов процессора, для виртуальных памятей — сотни тактов, они очень затратны по энергетике, приблизительно в 300 раз больше, чем к внутренней памяти и 600-700 раз затратнее, чем выполнение одной операции над числами FP64. По этой причине важно наличие большого объема внутрикристальной памяти, чтобы как больше обращений приходилось к ней. Если отношение пиковой производительности к объему внешней памяти говорит о достаточности ее для хранения данных для вычислений на таких скоростях, то соотношение внутрикристальной памяти и внекристальной говорит о возможной доле эффективного выполнения обращений к памяти. В этом качественный смысл еще одного показателя сбалансированности, приведенного ниже.

1. *Показатель сбалансированности* **ΣVHBM/DDR4/ΣVСASH/SMEM.** Чем этот показатель меньше, тем лучше, поскольку он отражает долю обращений к памяти (его обратная величина), которая при равновероятных условиях может попасть во внутрикристальную память. Для GPU Volta этот показатель 1000, а для GPU Ampere он уже равен 490. В этом процессоре значительно увеличена кэш-память уровня L2, а на самом деле этот показатель равен еще меньшему значению (около 2-х раз), поскольку усложнилась логика работы L2, например, в нем применяется компрессия хранимых данных.

Для платы M2000 этот показатель равен 67, поскольку для процессоров Colossus перенос основной работы на внутрикристальную память рекордно большого объема — принципиальное решение.

В процессоре NM6408MP объем внутрикристальной памяти порядка 10 MB (т.е. 0.01 GB). По этой причине данный показатель сбалансированности равен 5GB/0,01GB = 500. Формально это на уровне GPU Ampere, в действительности, может, в 2 раза хуже. Это не так плохо, но лучше бы было меньше. Это решается в проекте mPX[12].

# **Варианты суперкомпьютерных узлов на базе NM6408MP**

По оценкам инженеров, минимальный размер отдельной платы с одним NM6408MP и 5 GB памяти — это 12х12 см. Если взять стандартную серверную плату размером 48х89 см, то по ширине может разместиться 4 таких однопроцессорных платы, а в длину (если им отвести 2/3 этой длины, остальное на питание и охлаждение, интерфейсы), то разместится 5 однопроцессорных плат. Итого, в самом лучшем случае разместится 20 однопроцессорных плат. На самом деле, наиболее реально размещение 8 плат, максимум 12—16.

На рисунке (см. Рисунок 8(а)) представлен вариант 8-процессорной платы с прямыми соединениями по EL-линкам. Каждый процессор доступен по PCIe интерфейсу. Применяются два коммутатора PCIe. Межузловая сеть — отечественная сеть Ангара, подключена к хост-процессору.

На рисунке (см. Рисунок 8(б)) представлен вариант 16-процессорной серверной платы. Он мало отличается от варианта 8-процессорной платы.

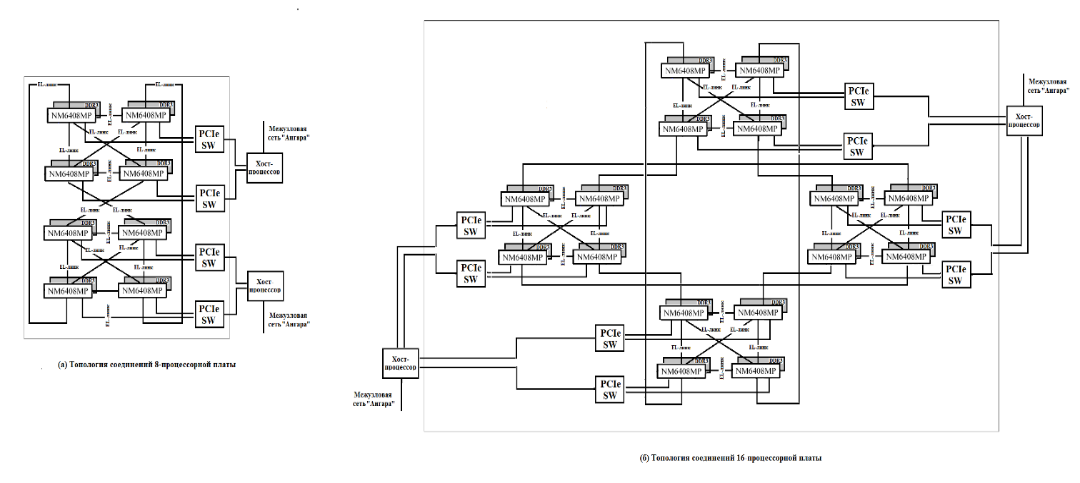


Рисунок 8. Топология 8 и 16-процессорной серверной платы

Оба представленных варианта могут работать с программным обеспечением БУПВ/МХ, в котором с EL-линками возможна работа как в мультиплексном, так и селекторном режиме (см. далее). Представленный на Рисунок 9 вариант ориентированы только на работу с EL-линками в мультиплексном режиме. Это очень общий вариант реализации платы, в них NM6408MP выступают как вычислители и как маршрутизаторы внутриплатной сети.

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\final\рис9_1.tif

Рисунок 9. Серверные платы с соединениями 2D-тор и через общий коммутатор.

Замечено, что в зарубежных многопроцессорных платах с GPU, Gaudi и Сolossus на платах со временем появляются блоки центральной коммутации линков. В платах DGX-2 и DGX — это несколько коммутаторов NVSwitch, в плате HLS-1 с Gaudi это Ethernet-коммутатор “каждый с каждым”, в плате M2000 это решено наиболее изящно ­— реализована новая микросхема-мост IPU-gateaway, которая не только коммутирует IPU-линки, но и обеспечивает процессорам общий интерфейс с памятью обмена в 480 GB с пропускной способностью в 180 TB/s. В платах с NM6408MP такого пока нет, но возможен вариант, схожий с изображенным на Рисунок *9*10.

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\final\рис10.tif

Рисунок 10. Узел с 6-рооцессорной платой с центральным коммутатором

# **Проблемы реализации БУПВ/MX для мультипроцессорных систем и варианты их решений**

В предыдущих разделах были рассмотрены разные варианты зарубежных многопроцессорных плат с процессорами-ускорителями, а также варианты таких плат с процессорами-ускорителями NM6408MP. Исходя из баланса производительностей таких плат и хост-процессоров доказана целесообразность разработки таких отечественных плат даже при сильном превосходстве по производительности зарубежных процессоров-ускорителей.

В настоящее время имеется готовый вариант 5-процессорной платы (см. Рисунок 1). Этот вариант и другие возможные в будущем варианты (см. Рисунок 8, 9, 10) можно уже сейчас программировать с использованием библиотек нижнего уровня [3], а также с использованием высокоуровневой библиотеки БУПВ/М [3,4] применительно к каждому процессору платы в отдельности, а межпроцессорные взаимодействия – с использованием библиотек нижнего уровня. Единственное условие такого использования – наличие доступа от хост-процессора к каждому процессору-ускорителю платы через интерфейс PCIe.

Вместе с тем, актуален вопрос разработки более удобных средств параллельного программирования в виде библиотеки БУПВ/МХ, продолжающей заложенную в БУПА/М концепцию применения общепринятых моделей параллельного программирования и сокрытия от пользователей особенностей архитектуры процессоров-ускорителей, но не теряя при этом преимуществ от использования этих особенностей.

Библиотека БУПВ/М поддерживает “плоскую” модель процессов MPI, а также одноуровневую модель с разделяемой общей памятью PGAS. Ставится цель это реализовать и в БУПВ/МХ, однако главной проблемой тут является отсутствие возможности работы с ресурсами одного процессора (регистры и память) из другого через общее адресное пространство. В одном процессоре NM6408MP это было возможно. Для преодоления такой проблемы необходимо решить следующие две первоочередные задачи.

1. Реализовать схемы взаимодействия процессов через внекристальные EL-линки.
2. Программно реализовать (эмулировать) выполнение операций с ресурсами процессоров платы через общее логическое адресное пространство с использованием интерфейcов PCIe. Какой бы неприемлемой не показалась эта идея из-за предполагаемой неэффективности, но такая работа с PCIe производилась в других проектах.

Далее описываются варианты решения этих задач, более подробно для EL-линков, однако опережая эти описания, отметим, что с решением поставленных задач связано введение следующих скрытых от пользователя системных транзитных процессов, выполняемых на соответствующих узлах в режиме разделения времени.

1. На уровне ARM-узлов кластеров вводятся три процесса, работающих в разных режимах использования EL-линков – MX, MS и MR.
2. На уровне центрального ARM-узла вводится процесс MY, который участвует в реализации операций с памятью через общее логическое адресное пространство, покрывающее все ресурсы процессоров платы.

# **Организация работы с EL-линками**

NM6408MP имеет 4 EL-линка. По каждому из них передача сообщений может происходить в одном и другом направлении с пропускной способностью для направления до 1.7 GB/s (теоретически достижимая 2 GB/s).

Каждый из четырех EL-линков подключен к ARM-узлу соответствующего кластера через 3-й порт 4-х портового коммутатора, принадлежащего этому узлу (cм. LCхх на Рисунок 11). Порт 2 этого коммутатора соединен с компортом CP0 этого узла, только через этот компорт возможно взаимодействие с процессами и памятью этого узла. Если это требуется для EL-линка, то в коммутаторе должно быть установлено соединение портов 3 и 2. На рисунке (см. Рисунок 111(a))) проиллюстрировано такое соединение для обоих направлений передачи сообщений по линку, это используется для реализации схемы мультиплексорного режима использования EL-линка (см. Рисунок *11*12).

Передача через EL-линк возможна и через порты 0 или 1. При этом поскольку все каналы передачи дуплексные, то и EL-линк может передавать и принимать сообщения одновременно по разным портам, это зависит от настройки 4-х портового коммутатора. На рисунке (см. Рисунок 11(б)) проиллюстрированы варианты такого соединения, это используется для реализации схемы селекторного режима применения EL-линка, что описывается далее (см. Рисунок 13). Порт 1 всегда подключен к компорту CP0 0-го NM-ядра кластера, а порт 0 к соответствующему кластеру компорту CCP(i) центрального узла процессора. Это также отображено на рисунке (см. Рисунок *11*(b)).

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\рис11.tif

Рисунок 11. Две схемы использования EL-линка, три типа транзитных процессов

Определенные затруднения в использовании рассмотренных возможностей коммутации связаны с тем, что их надо задавать пользователю установкой управляющих регистров. Именно это для внутрикристальных маршрутов скрывает от пользователя библиотека БУПВ/М, а для EL-линков это предстоит реализовать в БУПВ/МХ. Главная проблема здесь в отсутствии прямого доступа к управляющим регистрам одного процессора из другого. Используя рисунок (см. Рисунок 11), это можно конкретизировать так – как при передаче из процессора B в процессор А настроить из процессора B в процессоре А коммутатор LCxx и компорт СP0 или CP-компорт 0-го NM-узла кластера или центрального узла. Рассмотрим два варианта решений, они проиллюстрированы на рисунке (см. Рисунок 11).

**Вариант “Двунаправленный мультиплексный режим использования EL-линка”.**

Это самое общее решение, позволяющее производить любые операции из одного процессора в другом, но не обеспечивает эффективное использование EL-линка для передачи сообщения. Суть решения в том, что при загрузке программы коммуникационные ресурсы заранее получают постоянную настройку, которая предполагает работу с ними по фиксированному протоколу транзитных процессов с одной и другой стороны EL-линка. Схема такого решения для процессоров “A” и “B”, взаимодействующих через EL-линк, показана на Рисунок 12. Организацию взаимодействия через EL-линк по приведенной схеме кратко опишем в виде следующих семи пунктов.

1. Коммутаторы по одну и другую сторону EL-линка настроены на соединение портов 3 и 2 по двум направлениям. При этом это не исключает возможность взаимодействия портов 0 и 1, они могут быть соединены или нет.
2. Компорты CP0 по одну и другую сторону EL-линка настроены на передачу сообщений фиксированной длины L\_buf (см. Рисунок 12). Поскольку в процессе передачи сообщения значения управляющих регистров изменяются, то после каждой передачи такого сообщения принятая стандартная настройка программой PMX процесса MX восстанавливается. Об окончании выдачи сообщения процесс PMX оповещается прерыванием OK-send от компорта CP0. Об окончании приема сообщения процесс PMX оповещается прерыванием OK-recv, также от CP0.

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\final\рис12.tif

Рисунок 12. Схема организации в одном CPU кластера пересылок сообщений через транзитный процесс MX

1. В передаваемом сообщении длины L\_BUF содержательная часть — пакет (cм. Pkt на рис.12). Pkt имеет заголовок (head), тело (body) и хвост (tail). Он может быть двух типов — управляющий (Pktcontrol) и информационный (Pktinf). Pktcontrol — это команды разного типа, которые выдает один процессор другому (операции с памятью, операции с сигналами или событиями, операции установки маршрутов или барьерной синхронизации и др.). Pktinf используются при передаче сообщений. Тип пакета и код операции указывается в его заголовке.
2. Транзитные процессы MX по обе стороны EL-линка работают на соответствующих кластерных ARM-узлах асинхронно, в режиме разделения времени ARM-узла совместно с процессами параллельной программы, соответственно PСPU-A и PСPU-B.

Все процессы параллельной программы передают информацию в соответствующий MX для формирования Pkt через буфер FIFOSEND-OUT, а принимают информацию от MX либо напрямую, если это управляющий пакет Pktcontrol (условно обозначено точкой F), либо для информационных пакетов Pktinf через буфера QUEUERECV и QUEUESEND-IN (условно обозначено точкой G). Эти буфера предполагается разместить в подключенной соответствующему кластеру DDR-памяти.

1. В зависимости от длины передаваемого сообщения, используются три протокола их передачи. Выделяются короткие сообщения (SHORT, 64 байта), средние (EAGER, около килобайта), а также длинные (LONG, произвольная длина). Команды имеют вид короткого сообщения. Размер буферов непосредственного приема/выдачи (L\_buf), а также элементов буферов FIFOSEND-OUT, QUEUERECV и QUEUESEND-IN равны длине короткого сообщения.
2. Процесс выдачи (для примера) короткого сообщения состоит в следующем. Сначала для выполняемой функции БУПВ/МХ определяется EL-линк, в который следует передать короткое сообщение. Далее выполняются следующие два этапа.
   1. Этап 1 — пересылка сообщения из процесса пользователя в FIFOSEND-OUT кластера, где расположен EL-линк.
   2. Этап 2 — процесс MX выбирает очередной элемент из FIFOSEND-OUT (точка A1 на Рисунок 2) далее записывает его при появлении возможности в свой BUF-out (точка A2) и запускает выдачу его компортом CP0 через EL-линк. После получения OK-send этот процесс будет опять готов выдать короткое сообщение.
3. Процесс приема (для примера) короткого сообщения выглядит, в свою очередь, так.
   1. Этап 1 — получение от своего CP0 прерывания OK-recv, которое говорит о полученном в BUF-in сообщении. Прием этого сообщения процессом MX (точка A3).
   2. Этап 2 — анализируется заголовок принятого пакета. Если это Pktcontrol, то он дешифрируется и выполняется имеющаяся в нем операция (точка A4). Если это Pktinf, то он отправляется в QUEUESEND-IN (точка A5) и далее рассматривается QUEUERECV, на предмет встречи пришедшего сообщения с запросом на его получение (“рандеву”), которое, возможно, было выдано ранее.

Если в QUEUERECV есть запрос на прием пришедшего сообщения, то организуется передача этого сообщения в выдавший запрос процесс от имени процесса MX. Для этого предполагается использовать блоки асинхронной пересылки MDMAC.

Если запроса в QUEUERECV нет, то Pktinf просто оставляется в QUEUESEND-IN и ожидает в нем запроса в QUEUERECV. Запрос в QUEUERECV появляется при выполнении в каком-либо процессе программы функции типа Receive. После постановки такого запроса в QUEUERECV также проверяется ситуация типа “рандеву”, т.е. нет ли уже в QUEUESEND-IN требуемого сообщения.

**Вариант “Односторонний селекторный режим использования EL-линка”**

Если просто требуется передавать сообщения, то можно использовать селекторный режим работы с EL-линком. Этот режим поясним на примере схемы пересылки из процесса “B” в процесс “A”, представленной на рисунке (см. Рисунок 13). В целом, это похоже на логику работы статических маршрутов, реализованных в БУПВ/М.

1. Направление EL-линка от процессора “B” к “A” (обозначено “C”) используется только для передачи сообщений. Процессом-поставщиком может быть процесс на кластерном узле CPU (PM), на центральном узле ССPU, на 0-м NM-узле этого кластера. Для связи с этим процессом в 4-х портовом коммутаторе LCхх должно быть установлено соединение порта 3 с портом 2, либо 0, либо 1. Процессом-получателем может быть также процесс на кластерном узле CPU (PN), на центральном узле CCPU, на 0-м узле кластера. Для соединения с ним в 4-х портовом коммутаторе LCxx процессоре “A” должно быть установлено соединение порта 3 с портом 2, 0 или 1.
2. Направление EL-линка от процессора “A” к процессору “B” (обозначено OKSR) предназначено для передачи служебных сообщений, которыми могут быть либо подтверждения о завершении приема сообщения, либо специальные команды (см. далее). В процессоре “A” пересылка такого сообщения производится транзитным процессом MS, а в процессоре “B” сообщения получает транзитный процесс MR. Логика работы этих процессов такова.
   1. Процесс MS выбирает сообщения из буфера FIFOOK-out (точка A1). Этими сообщениями могут быть уведомления об окончании приема по C-направлению EL-линка в процессе получателе. Такие уведомления формируются после окончания приема в компорте этих процессов. Дополнительно могут передаваться специальные команды. Например, такой командой может быть команда перекоммутации соединения без выхода из селекторного режима. Выбранное из FIFOOK-out сообщение передается через OK-buf в OKSR EL-линка через компорт CP0 (точки A2 и A3). Готовность к новой такой передаче восстанавливается по сигналу OK-send конца выдачи от компорта CP0.

2.2. Процесс MR принимает сообщения по направлению OKSR напрямую через буфер OK-buf, сигнал о наличии сообщения - OK-recv от компорта CP0 (точки B1 и B2). Далее это сообщение помещается в буфер FIFOOK-in (точка B3). MR также выбирает сообщения из FIFOOK-in (точка B4), далее, в зависимости от типа сообщения, передает уведомление в процесс-поставщик, либо выполняет специальную команду из этого сообщения.

C:\Users\User\Desktop\конференция\Микроэлектроника 2020\картинки\рис13.tif

Рисунок 13. Схема организации пересылки сообщения в селекторном режиме с использованием транзитных процессов MR и MS.

3. Переход в селекторный режим происходит по специальным командам, которыми процессы MX, работающие с EL-линком обмениваются. При инициализации параллельной программы устанавливается мультиплексный режим.

4. Возврат из селекторного режима в мультиплексный может происходить по инициативе принимающей стороны (процессора “A”) посредством выдачи по OKSR специальной команды, как это описано в разделе 2.1.

5. Для контроля состояния (режима использования) EL-линка на обоих его сторонах имеется специальная структура данных, называемая драйвером EL-линка.

Таким образом, для обеспечения работы в мультиплексном и селекторном режимах в каждом кластере должны быть три транзитных процесса — MX, MS и MR, всего в процессоре будет 12 таких процессов.

# **3.2 Организация работы с физической памятью через общее адресное пространство**

Обеспечение такой возможности в вычислительных системах — одна из главных тем в последние 20 лет. Эффективная реализация этого на уровне аппаратных средств ожидается в будущих инновационных суперкомпьютерах экзафлопсного уровня в середине текущего десятилетия, в которых и суммарная физическая память вычислительных узлов будет исчисляться экзабайтами.

В библиотеке БУПВ/МХ такая возможность реализуется программно. Базовая концепция заключается в том, чтобы такая возможность была, а эффективность ее реализации в данный момент — это вторичное.

Логика реализации достаточно проста и прямолинейна.

В узлах CCPU вводится системный транзитный процесс MY, который через буфер получает в виде коротких сообщений заявки на выполнение операций над памятью через общее адресное пространство и далее передает их в хост-процессор. В хост-процессоре также вводится новый транзитный процесс MY, он выбирает поступающие в его буфер заявки на выполнение операций и распределяет их по процессорам в соответствии с имеющимся у него отображением логического адресного пространства на физические памяти процессоров.

Процессы MY в процессорах получают заявки на выполнение операций и выполняют их. После этого выдают подтверждение о выполнении операций. Далее эти подтверждения передаются процессом MY хост-процессора в выдавший заявку на выполнение процесс MY некоторого процессора.

Приведенная схема типична при эмуляции команд работы с памятью через общее адресное пространство. Для повышения пропускной способности выполнения этих команд в продвинутых реализациях применяется эмуляция мультитредовости ядер процессоров. В библиотеке БУПВ/МХ применение этого пока не планируется.

# **Выводы**

Разработка БУПВ/М доказала состоятельность идеи эффективной реализации модели параллельных программ, близкой к MPI и PGAS на специализированном NM6408MP без потери его преимуществ. Разработка библиотеки БУПВ/МХ для многопроцессорных плат с NM6408MP продвигает базовую идею БУПВ/М.

Опыт проведенных работ позволяют сделать некоторые обобщения и сформулировать ряд предложений по развитию архитектуры процессоров семейства NeuroMatrix. Выделим предложения на ближайшую и дальнейшую перспективу, этот список формулируется впервые, в последующем он будут уточняться.

Ближайшая перспектива : увеличение количества мониторов в 0-х блоках локальной памяти узлов ARM-node; введение в систему команд узлов NM-node команд барьерной синхронизации обращений к памяти; введение “ждущих” состояний процессорных ядер; исключение X-маршрутов введением коммутатора компортов в узлах ARM-node; увеличение количества линков хотя бы до десяти, повышение их пропускной способности, расширение их функциональности добавлением RDMA (удаленного обращения к памяти через передаваемое сообщение).

Дальняя перспектива: мультитредовая организация ядер; разнотипные специальные функциональные устройства; использование локальной памяти с большим расслоением; переменная разрядность адресации памяти; атомарные операции для любого слова памяти любого уровня; введение почтовых ящиков и средств передачи сигналов; введение операций с битами слов памяти; разработка единой внутрикристальной сети; аппаратно-программная реализация операций с глобально адресуемой памятью [12, 14].

# **Заключение**

Авторы считают, что в данной работе новыми являются следующие положения и результаты:

1. Систематизированы данные по серверным узлам суперкомпьютеров на базе многопроцессорных плат с процессорами-ускорителями GPU Volta, GPU Ampere, Gaudi и Colossus, проведен анализ их структур и вычислены показатели их сбалансированности.
2. На основе показателей сбалансированности многопроцессорных плат зарубежного производства, доказана целесообразность разработки 8-ми и 16-процессорных серверных плат с процессором NM6408MP для узлов суперкомпьютеров с отечественными универсальными процессорами производительности уровня Эльбрус 8СВ.
3. Сформулированы варианты решения первоочередных задач разработки БУПВ/МХ — организация мультиплексного и селекторного режима использования межпроцессорных EL-линков, эмуляция операций с физической памятью процессоров платы через логическое глобальное адресное пространство.
4. Исходя из опыта разработки БУПВ/М и подготовки проекта разработки БУПВ/МХ сделан ряд предложений по развитию архитектуры процессоров семейства NeuroMatrix в ближайшее время и на перспективу.

Сотрудники ЗАО “НТЦ” Модуль” одними из первых получили опыт разработки нейропроцессора в виде NM6408MP, а также его прикладного и системного программного обеспечения, теперь приступила к разработке многопроцессорных плат и их программного обеспечения для специализированных суперкомпьютеров. Полагаем, что любые подробности по выполненным и запланированным работам будут полезны коллегам.

Данная статья является сокращенным вариантом работы [14], предоставляемой по запросу.

**Литература**

[1] Эйсымонт А.Л., Черников В.М., Черников Ан.В., Черников Ал.В., Косоруков Д.Е., Насонов И.И., Комлев А.А. Гетерогенная многопроцессорная система на кристалле с производительностью 512 Gflops // Системы высокой доступности, 2018, т.14, №3, с.49-56.

[2] Модуль NM Card // https://www.module.ru/products/2-moduli/nm-card

[3] Биконов Д.В., Пузиков А.Д., Сивцов А.С., Эйсымонт Л.К. Трехуровневая система параллельного программирования гибридного 21-ядерного скалярно-векторного микропроцессора NM6408MP // Вопросы кибербезопасности, номер 4, 2019, стр.22-34.

[4] Биконов Д.В., Бражкин А.А., Пузиков А.Д., Сивцов А.С., Дементьев В.В., Спесивцев Д.А., Эйсымонт Л.К. Высокоуровневая система параллельного программирования многоядерного гибридного процессора // Труды IV Международного Форума “Микроэлектроника 2020” (28 сентября – 03 октября 2020, г.Ялта), 10 стр.

[5] NVIDIA Tesla V100 GPU Architecture. The World’s Most Advanced Data Center GPU // August 2017, 58 pp.

[6] Kennedy P. NVIDIA DGX-2 Details at Hot Chips 30

[7] NVIDIA A100 Tensor Core GPU Architecture. Unprecedented acceleration at every scale // May 2020, 83 pp.

[8] Gwennap L. Habana Offers Gaudi for AI Training //Microprocessor report, June 17, 2019, 4 pp.

[9] Gaudi Training Platform White Paper // Habana Labs, June 2019, 25 pp.

[10] Toon N. Introducing 2nd generation IPU systems for AI at scale // //15 july 2020.

[11] Jia Z., Tillman B., Maggioni M., Scarpazza D.P. Dissecting the Graphcore IPU Architecture via Microbenchmarking.// Technical report, High Performance Computing R&D Team Citadel, 7 Dec 2019 - 91 pp.

[12] Эйсымонт Л.К., Никитин А.И., Биконов Д.В., Бражкин А.А., Пеплов И.C., Федоренко П.П., Ермаков С.С., Эйсымонт А.Л., Комлев А.А. Исследовательский проект массово-параллельного процессора на базе мультитредовых ядер со специализированными ускорителями // Вопросы кибербезопасности, номер 3, 2020, стр.39-55.

[13] https://www.graphcore.ai/benchmarks

[14] Биконов Д.В., Эйсымонт Л.К. и др. Многопроцессорные платы на базе NM6408MP и их системное программное обеспечение. Технический отчет 02-2020, ЗАО”НТЦ”Модуль”, 2020, 23 стр.