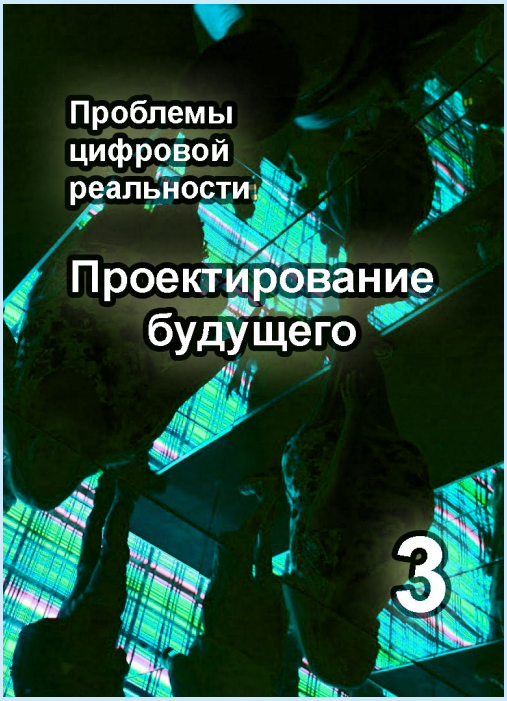
ИПМ им.М.В.Келдыша РАН

Будущее-2020 • Труды конференции



А.А. Адамов, Л.К. Эйсымонт

Варианты архитектурных решений ЭКБ для систем искусственного интеллекта

Рекомендуемая форма библиографической ссылки

Адамов А.А. Эйсымонт Л.К. Варианты архитектурных решений ЭКБ для систем искусственного интеллекта // Проектирование будущего. Проблемы цифровой реальности: труды 3-й Международной конференции (6-7 февраля 2020 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2020. — С. 112-131. — https://keldysh.ru/future/2020/10.pdf https://doi.org/10.20948/future-2020-10

Размещено также видео выступления

Варианты архитектурных решений ЭКБ для систем искусственного интеллекта

А.А. Адамов, Л.К. Эйсымонт

Закрытое акционерное общество «НТЦ “Модуль”»

Аннотация. Взрывной рост количества создаваемых процессоров для работы с нейросетями (нейропроцессоров) и других образцов электронно-компонентной базы (ЭКБ) для систем искусственного интеллекта (ИИ) предлагается рассматривать как новое направление развития ЭКБ суперкомпьютерных технологий (СКТ), для которого еще в большей степени важно преодоление проблем по работе с памятью, энергоэффективность, повышение параллелизма операций и снижение накладных расходов на коммуникации. Делается предостережение по поводу того, что в работах по технологиям ИИ в целом, необходимо учесть ошибки организации работ по отечественным СКТ, поскольку в этом 20-летнем процессе был допущен перекос – наибольшее внимание уделялось всему кроме развития отечественной ЭКБ. В результате образовалось сильнейшее отставание от зарубежной ЭКБ СКТ и, как следствие, возникли значительные современные проблемы импортозамещения в этой области. Работы по отечественным технологиям ИИ, также как по технологиям СКТ, могут быть разделены на два направления: первое – адаптация и развитие зарубежных технологий, разработка оригинальных алгоритмов и программного обеспечения; второе – разработка собственной ЭКБ ИИ. При этом работы по ЭКБ ИИ, по мнению авторов, имеют большую значимость, что объясняется будущей массовостью и разнообразием рынка систем ИИ, но при прогнозируемой остановке развития микроэлектронных КМОП-технологий (комплементарная структура металл–оксид–полупроводник), в то время, как промышленное освоение новых («пост-Муровских») технологий ожидается лишь в конце предстоящего десятилетия. При освоении отечественных СКТ, особую роль сыграли целевые программы Союзного государства (ЦПСГ) СКИФ и ТРИАДА. Для становления отечественных технологий ИИ также требуется, как минимум, российская ФЦП, а лучше ЦПСГ. Создаваемые в рамках таких программ образцы ЭКБ должны поддерживать не только нейронные сети глубокого обучения (DNN) и нейроморфные (спайковые) сети (SNN), но и другие базовые методы ИИ.

В статье с учетом истории развития и современного состояния работ по отечественным СКТ рассматриваются вопросы организации программ по технологиям ИИ с акцентом на разработку ЭКБ ИИ.

Ключевые слова: электронно-компонентная база, нейропроцессоры, мультитредовое ядро, проблемно-ориентированная архитектура,

нейронные сети глубокого обучения, нейроморфные сети, энергоэффективность

Variants of hardware architectural solutions for artificial intelligence systems

A.A. Adamov, L.K. Eisymont

Research Center "Module"

Abstract. The explosive growth in the number of created processors for working with neural networks (neuroprocessors) and other samples of the electronic component base (ECB) for artificial intelligence systems (AI) is proposed to be considered as a new direction in the development of ECB supercomputer technologies (HPC tech), for which an even greater degree. It is important to overcome the problems of working with memory, energy efficiency, increasing the parallelism of operations and reducing the overhead of communication. A warning is made that in the work on AI technologies in general, it is necessary to take into account the errors in organizing work on domestic HPC, since skewing was allowed in this 20-year process – most attention was paid to everything except the development of the domestic electronic components . As a result, a severe lag formed behind the foreign ECB HPC and, as a result, there were significant modern problems of import substitution in this area. Work on domestic AI technologies, as well as on HPC technologies, can be divided into two areas: the first is the adaptation and development of foreign technologies, the development of original algorithms and software; the second is the development of its own electronic components. At the same time, the work on electronic components of AI, according to the authors, is of great importance, which is explained by the future mass character and variety of the market of AI systems, but with the predicted halt in the development of microelectronic CMOS (complementary metal-oxide-semiconductor) technologies, while the industrial development of new ones ("post Moore") technology is expected only at the end of the coming decade. In the development of domestic HPC, a special role was played by the target programs of the Union State (CPSG) SKIF and TRIAD. For the establishment of domestic AI technologies, at least a Russian federal target program, and preferably a CPSG, is also required. ECB samples created in the framework of such programs should support not only deep learning neural networks (DNN) and neuromorphic (spike) networks (SNN), but also other basic AI methods.

Taking into account the history of the organization and development of work on domestic HPC technology, as well as their current state, the article discusses the organization of AI technology programs with a focus on the development of electronic components of AI.

Keywords: AI hardware, neuroprocessors, manytiled architecture, multithreaded core, domain specific architecture, deep learning neuro networks, neurumorthic (spaiking) networks, training, inference

1. Предыстория и проблемы отечественной ЭКБ СКТ

Созданием ЭКБ для ИИ, но в виде специализированных процессоров и компьютеров, занимались еще много лет назад [1,2]. Тогда темы реализации нейронных сетей еще не было, а упомянутые процессоры были ориентированы на реализацию языков Лисп и Пролог [1], в нашей стране – на отечественный язык Рефал в виде, так называемого, символьного процессора. Отличительным свойством этих процессоров была эффективная по быстродействию реализация операций со списковой памятью.

Лисп- и Пролог-процессоры были реализованы, но вопреки большим ожиданиям 1980-х гг., не выдержали конкуренции с появившимися универсальными микропроцессорами. Символьный процессор был реализован микропрограммно, эта реализация также оказалась неперспективной.

Только в настоящее время тема специализированных процессоров для работы со списковой памятью возродилась в процессорах для работы с графами [3,4,5] (это направление 2 из 5 главных проблем по ЭКБ СКТ, выделенных в [6]).

Работы по спецпроцессору под язык Рефал были начаты в первой половине 1970-х гг. в рамках инициированных в ИПМ им. М.В. Келдыша АН СССР фундаментальных исследований по архитектуре суперкомпьютеров и высокопроизводительных систем (СК и ВС), обоснование исследований сделано в работе [7]. Возникновение этих работ было естественным шагом, поскольку практический опыт предыдущих проектов давал возможность понять, как должны быть устроены в то время «правильные» СК и ВС.

Это направление исследований кроме теоретических результатов и получения практических навыков имитационного моделирования СК и ВС имело еще один результат – образование архитектурной школы ИПМ. Профессиональная подготовка и просто воспитание специалистов в этой школе, давно вышедшей за границы ИПМ и насчитывающей сейчас около четырех поколений, позволили её воспитанникам участвовать во многих проектах, связанных с разработкой и эксплуатацией СК и ВС.

Такой исторический опыт важен тем, что при организации работ по освоению технологий ИИ, должны быть также сформированы стабильные высокопрофессиональные коллективы в общей сложности в несколько сотен сотрудников, которым предстоит работать, как минимум, до конца предстоящего десятилетия. Приведем примеры затрат ресурсов на выполнение зарубежных разработок по ЭКБ ИИ.

Пример 1 – система на кристалле Xavier фирмы NVIDIA для беспилотных транспортных средств разрабатывалась коллективом в 2000 человек в течение 4 лет, было затрачено около 2-х миллиардов долларов.

Пример 2 – стартаповская израильская фирма Habana Labs в 150 человек разработала два успешных микропроцессора для области ИИ – Gaudi (обучение сетей типа DNN) и Goya (вывод решений с использованием сетей DNN). Фирма получила поддержку в $250 млн от Intel, а в декабре 2019 г. вообще была куплена фирмой Intel за 2 миллиарда долларов. Вот пример селекции специалистов по-американски, а до этого Intel купила другую израильскую фирму – Nervana с их изделиями, но потом это направление частично закрыло. Вот уж действительно, процесс селекции не прост и жесток.

Далее для изучения опыта предыдущих проектов перейдем ко времени формирования современных СКТ. Этот процесс был связан с лавинообразным появлением и сразу внедрением микропроцессоров. Может показаться неожиданным, но в экспертной среде это время считается временем «темного средневековья» для исследований по архитектурам компьютеров, поскольку особо думать не приходилось, просто надо было ставить побольше микропроцессоров в СК и ВС, а эти микропроцессоры появлялись волнами, все лучше и лучше со сменой микроэлектронных технологий.

Это было «золотое время» развития вычислительных средств за счет развития микроэлектронных технологий в соответствии с законом Мура (удвоение плотности размещения транзисторов каждые 18 месяцев) и законом Денара (пропорциональное улучшению технологий сокращение потребляемой энергии транзистором, что не позволяло расти энергопотреблению кристаллов с более плотной компоновкой).

Венцом того бурного этапа развития стало появление в 1990-х гг. технологий сборки СК и ВС в виде кластеров из готовых и коммерчески доступных компонентов (серверные платы, процессоры, память), коммуникационное оборудование (маршрутизаторы, коммутаторы, кабели), стандартное программное обеспечение (операционные системы, компиляторы, инструментальные средства, системные и прикладные библиотеки). В то время и начали использовать термин технологии СКТ.

По существу, это технологии «отверточной сборки», но их освоение оказалось также не простым делом. Первыми в нашей стране их освоили специалисты ВНИИЭФ (Всероссийский научно-исследовательский институт экспериментальной физики, г. Саров) и ФГУП «НИИ “Квант”» для создания уникальных суперкомпьютеров, применяемых при решении важных государственных задач.

Широкое освоение этих технологий в России и Беларуси началось в 2000 г. в рамках программ Союзного государства СКИФ (головное предприятие от России – Институт программных систем РАН, особая заслуга в этом Альфреда Карловича Айламазяна, решающее участие в организации проекта принял также НИЦЭВТ (Научно-исследовательский центр электронной вычислительной техники) [8,9,10]) и ТРИАДА

(головное предприятие от России – НИЦЭВТ, один из авторов данной статьи лично составил эту программу и ее обоснование, но заслуги в ее отстаивании в российских госструктурах в большей степени принадлежат А.И. Слуцкину и В.В. Митрофанову).

Появление в широком использовании зарубежной кластерной технологии создания СК и ВС способствовало разработке приложений, связанных с высокопроизводительными вычислениями (HPC). Все это вызвало некоторую «эйфорию» от успехов, особенно в наших странах, поскольку в условиях существовавшего тогда развала научно-технической сферы удалось по HPC без особых усилий встать в ряд передовых зарубежных стран.

Кластерные технологии используются в мире и сейчас, достаточно изучить типы суперкомпьютеров, включенных в рейтинговый список Top500. Но для части разработчиков СК и ВС такое головокружение от успехов продолжалось недолго.

В нашей стране сотрудников НИЦЭВТ еще летом 1999 г. заинтересовало сильное расхождение объявляемой пиковой производительности кластеров и реально развиваемой. Это было замечено на тестах НАСА из области аэрогидродинамики, которые при завершении выдавали сообщение о достигнутой на тесте реальной производительности. Оказалось, что она доходила иногда лишь до нескольких процентов от пиковой.

Изучение причин этого привело к пониманию следующих проблем [10], эти проблемы проявляются сейчас и при освоении технологий ИИ:

– наличия опасной «стены памяти», когда большие задержки выполнения обращений к памяти приводят к простою функциональных устройств из-за отсутствия данных для них;

– наличие совокупного влияния на масштабируемость производительности накладных расходов на коммуникации и нелинейного изменения производительности процессора из-за изменения эффективности использования кэш-памяти при дроблении из-за распараллеливания рабочего множества данных задачи на подмножества, приходящиеся на процессоры;

– наличие тенденции перспективных моделей параллельных вычислений к работе над общей памятью, что резко снижает потери на коммуникации.

Далее был выбран подход к решению этих проблем за счет использования процессоров с мультитредовой архитектурой [10,11], а также других архитектурных приемов. Такие работы появились уже в программе ТРИАДА, затем в нескольких отечественных НИР, а потом в проекте СКСН «Ангара» [12,13,14], который в целом не поддержали. Однако понимание перечисленных проблем и вариантов их решения

позволило понять значение появления многосокетных серверных узлов и освоить их эффективное использование [15,16], реализовать сеть «Ангара».

В то же время (2010 г.) появился прообраз современных GPGPU (а они сейчас проблема номер один в отечественных СКТ, это направление 1 из выделенных в [6] пяти направлений) именно как мультитредовый микропроцессор в образе NVIDIA GPU Fermi [17]. Фирма NVIDIA оказалась наиболее прозорливой в то время и более удачливой, чем НИЦЭВТ.

Это произошло не случайно, глобальное видение проблемности кластерных технологий, которое не удалось привить в нашей стране, удалось объяснить в США, в этом помогло появление в 2004 г. японского векторного суперкомпьютера Earth Simulator, в котором не использовались рыночные микропроцессоры, но который на несколько лет сбросил США с первого места в мире по HPC вычислениям.

Проведенный американскими специалистами углубленный анализ причин этой потери лидерства в области HPC показал, что технология кластерной сборки из коммерчески доступных компонентов не совсем достаточна. В частности, при построении СК и ВС для взятия очередного для того времени рекордного барьера по реальной производительности в один PF (петафлопс, 10^{15} операций над 64-хразрядными числами с плавающей точкой – FP64). Такой вывод привел в США к ряду действий на государственном уровне, в том числе по возрождению глубоких архитектурных исследований и разработок, к восстановлению в американских университетах школ по архитектуре компьютеров.

Можно это все охарактеризовать так, что настал «момент истины» (в смысле работы [7]), а в обществе оказались востребованы результаты фундаментальных работ по архитектуре компьютеров и ЭКБ. История этого «прозрения» и спада эйфории от кластерных технологий описана в работе [18]. Здесь лишь напомним, что в США такое восстановление работ было подкреплено сначала Законом США 108-423 (ноябрь 2004 г., “Department of energy high end computing revitalization act of 2004”), после чего последовали федеральные и ведомственные программы, важным оказался проект DARPA HPCS [19], который, в конечном итоге, успешно выполнили победители всех этапов этой программы – фирмы Cray и IBM. Похожие программы быстро были организованы в Китае и Японии. Это привело к появлению более мощной ЭКБ СКТ, которая уже сейчас либо просто коммерчески недоступна, либо доступна с большими ограничениями.

К сожалению, судя по состоянию современных отечественных СКТ, можно сделать вывод, что эйфория от рыночных кластерных технологий у нас продолжается до сих пор. Действительно, наряду с тем, что получены значительные достижения по многим направлениям СКТ, прорыва по ЭКБ СКТ не получилось.

Отставание по ЭКБ СКТ, которое на протяжении почти десяти лет замалчивалось и за которое даже непонятно, кому отвечать, стало сейчас критическим для национальной безопасности. Если по организации работ все оставить так, как есть, то изменений в лучшую сторону не будет.

Начало постановки работ по технологиям ИИ уже говорит о том, что высока вероятность повторить эти же ошибки. Уже заметны тенденции получения быстрых результатов и «победных рапортов», вводящих в заблуждение руководство страны. По этой причине и была выше изложена история становления отечественных СКТ. В свое время это излагалось и высказывались предостережения, но тогда это не подействовало. Далее посмотрим на сегодняшний результат – сравним для краткости только по процессорным компонентам уровни зарубежной и отечественной ЭКБ СКТ.

Современный уровень зарубежной ЭКБ СКТ можно оценить, по суперкомпьютеру SUMMIT Окриджской национальной лаборатории США (ORNL), он запущен 8 июня 2018 г.

Напомним, что ORNL и Аргонская национальная лаборатория (ANL) зафиксированы Законом 108-423 (ноябрь 2004) как «ультракомпьютерные центры» для решения задач особой важности, список которых утверждается лично Президентом США. Главный центр в ORNL, поэтому в экспертной среде с большим вниманием относятся к тому, что там происходит.

ORNL SUMMIT – это самый мощный суперкомпьютер в мире, его пиковая производительность 300 PF (FP64). Каждый вычислительный узел (ВУ) – это два микропроцессора IBM Power 9 (далее – Power 9), к которым подключены шесть плат с графическими процессорами NVIDIA GPGPU Volta (далее – GPU V100) через высокоскоростные линки NVlink (дуплексная пропускная способность 200 Gb/s у каждого).

GPU V100 и Power 9 –это два базовых компонента зарубежной ЭКБ в 2018 г. Главная функция Power 9 состоит в управлении вычислениями на GPU V100, что включает скалярные вычисления, в том числе адресные для доступа к большой DDR-памяти с целью подготовки данных для GPU и восприятия результатов.

В табл. 1 основные характеристики этих базовых компонентов приведены в сравнении с характеристиками отечественных микропроцессоров Эльбрус 16С и Эльбрус 8СВ, а также процессора обработки сигналов и нейровычислений NM6408MP разработки ЗАО «НТЦ “Модуль”». В этой таблице все показатели производительности пиковые, т.е. предельные, что могут дать все их функциональные устройств при идеальной загрузке.

Из табл. 1 видно, что заслуживающая глубокого уважения работа высокопрофессионального коллектива из МЦСТ и ИНЭУМ (Институт электронных управляющих машин) им. И.С. Брукса позволяет хоть и с

задержкой около 5 лет действительно приблизиться по пиковым характеристикам не только к процессорам Intel, что разработчиками обычно указывается, но даже к процессору Power 9. Сделаем лишь два замечания.

1. Процессор Power 9 в вариантах 24 и 12 ядер аппаратно поддерживает за счет динамического формирования сразу из выполняемых в ядре нескольких потоков команд пакет операций, выдаваемых за такт на функциональные устройства (это называется SMT-мультитредовость). Реализована одновременная работа с 96 потоками. Это при наличии большой пропускной способности интерфейса с памятью в 120 GB/s позволяет Power 9 быть толерантным (не снижать производительность) к возникающим большим задержкам выполнения обращений к внекристальной памяти в режимах нерегулярной работы с ней на некоторых задачах (это когда плохая пространственно-временная локализация адресов обращений), а также эффективно выполнять функции процессора-менеджера.

Таблица 1. Основные характеристики образцов зарубежной и отечественной ЭКБ СКТ

Характеристика	Микропроцессор (год серийного выпуска)				
	GPU Volta (2017)	Power 9 (2017)	Эльбрус 16С (2022)	Эльбрус 8СВ (2020)	NM6408 (2019)
Технология	12 нм	14 нм	16 нм	28 нм	28 нм
S – площадь кристалла	815 мм^2	695 мм^2	400 мм^2	350 мм^2	83 мм2
Количество транзисторов	$21{,}1\cdot10^9$	$8\cdot10^9$	$6\cdot10^9$	$3{,}5\cdot10^9$	$1{,}05\cdot10^9$
Частота	1,455 GHz	2,5-5 GHz	2 GHz	1,5 GHz	1 GHz
PW – потребление	300 W	190 W	100 W	90 W	25-30 W
P(FP64) – производительность	7,5 TF	0,435 TF	0,750 TF	0,288 TF	0,128 TF
P(FP32) – производительность	15 TF	0,870 TF	1,5 TF	0,576 TF	0,512 TF
P(FP16/FP32, DNN) – производительность	120 TF	—	—	—	—
Количество ядер	336	24, 12	16	8	21
Количество тредов	334064	96	16	8	21
BW(HBM/DDR) – пропускная способность памяти	800 GB/s (HBM)	120 GB/s	76 GB/s	68,2 GB/s	25,6 GB/s
BW(link) – дуплексная пропускная способность межкристальных линков	300 GB/s	300 GB/s	72 GB/s	16 GB/s	16 GB/s
P(FP32)/S – удельная производительность	18,405	1,252	3,75	1,646	6,169

(GF/мм^2)					
P(FP32, GF)/PW – энергоэффективность (GF/W)	50	4,58	7,5	6,4	17,07
BW(HBM/DDR) / P(FP32) – сбалансированность (B/F)	0,053	0,139	0,051	0,118	0,05

В процессорах Эльбрус формирование пакетов команд производится статически, на этапе подготовки программы компилятором, что снижает характеристики этих процессоров на задачах с нерегулярным доступом к памяти, это экспериментально было подтверждено на микробенчмарках и обсуждалось с разработчиками этих процессоров.

2. Большая пропускная способность межкристальных линков NVlink суммарно в 300 GB/s обеспечивает хорошие возможности построения вычислительных узлов с множеством Power 9 и процессоров-ускорителей. Вычислительный узел суперкомпьютера SUMMIT – один пример. Другой пример – 32-процессорный вычислительный узел суперкомпьютера Watson с иерархической сетью PERCS ограниченного диаметра (направление 3 работ, выделенных в [6]). Watson – мощнейший суперкомпьютер, применяемый фирмой IBM для решения задач ИИ.

Теперь перейдем к главному. Из табл. 1 видно, что основной компонент зарубежной ЭКБ, по которому образовалось отставание, – это ускоритель GPU V100. Так получилось, что именно этот компонент оказался на сегодняшний день из-за своей высокой производительности и доступности базовым и в области ЭКБ ИИ.

Наличие GPU в 2012 г. позволило впервые продемонстрировать на нейросети AlexNet уникальные возможности нейросетей в сравнении с обычными методами. Вычислительная сложность этого эксперимента составляла 0,01 петафлопс/день (когда машина с производительностью один PF работает полные сутки). Современные сети требуют уже затрат в 1000 петафлопс/день, причем это надо выполнять с малой задержкой, сеть должна обучаться за малое время.

Рассмотрим, чем же уникален GPU V100 по архитектуре?

GPU V100 содержит 84 потоковых мультипроцессора (SM-блоков), а каждый такой блок включает 4 мультитредовых ядра (MT-ядра). Высокая пиковая производительность в MT-ядрах обеспечивается наличием в них большого набора функциональных устройств, называемых CUDA-устройствами, это: 8 устройств для операций над числами FP64; 16 устройств для операций над числами FP32; 16 устройств целочисленных операций над 32-хразрядными целыми числами и двоичными кодами (INT16); 8 устройств вычисления адресов обращений к памяти; одно устройство выполнения специальных операций типа

математических функций; 2 устройства TPU выполнения тензорных операций над матрицами 4×4 из чисел формата FP16 и FP32. Для обеспечения плотной загрузки такого большого количества устройств в MT-ядре используются специальные архитектурные приемы – мультитредовая организация квазипараллельного процесса выполнения команд (32 асинхронных потока, WARP-а) и SIMT-принцип (синхронного выполнения потоков-тредов одного WARP-а). В GPU V100 запуск синхронных тредов был усовершенствован – теперь каждый из них имеет свой счетчик команд, раньше этот счетчик был только в WARP-е, а выдаваемый на CUDA-устройства за такт пакет операций выбирается из доступных WARP-у синхронных тредов, т.е. фактически реализована ограниченная SMT-мультитредовость. Один синхронный тред может работать с 16-ю 32-хразрядными регистрами. Запуск WARP-ов и тредов производится аппаратурой MT-ядер автоматически.

Такая архитектура GPU V100 позволила, например, по производительности на FP32, а это важный формат данных для ИИ, создать отставание в 10 раз главного и единственного отечественного образца процессора 2022 года от GPU 2017 г. Если же сравнивать производительности с учетом использования в GPU V100 тензорных устройств TPU (они поднимают производительность до 120 TF), то отставание почти в 100 раз.

В конце 2020 г. выйдет новый NVIDIA GPU Ampere, изготовленный по технологии 7 нм. Его архитектурные особенности пока неизвестны, но высока вероятность, что он будет похож на GPU Echelon, проект NVIDIA и Стэнфордского университета перспективного GPU, о котором стало известно в 2012 г., а последняя публикация с результатами проекта была в конце 2014 г. Этот проект также проанализирован в работе [20]. Здесь лишь отметим, что производительность GPU Echelon на FP64 объявлялась в 16 TF, т.е. уже в конце этого года возможно будет отставание более, чем в 20 раз? На нейровычислениях тогда следует ожидать отставание не менее, чем в 200 раз.

Вывод – в дополнение к Эльбрусу 16С или потом 32С нужен отечественный процессор-ускоритель, способный заменить NVIDIA GPU. Три возможных варианта решения этой задачи рассмотрены в [20].

Напомним, что первый процессор-ускоритель GPU Fermi появился в 2010 году, он содержал всего лишь 16 SM-блоков (двухядерных) и был изготовлен по технологии 40 нм, работал на частоте 0,5 GHz, имел производительность на FP32 1,33 TF, а на FP64 – 0,66 TF. Отечественный проект, который мог дать близкий результат – это процессор мультитредовой СКСН «Ангара», но в силу разных обстоятельств, этот проект не удался. Роковым для него также оказался 2010 г.

В итоге отечественного GPU нет. Отсюда еще один вывод – предлагаемые проекты с акцентом на развитие ЭКБ ИИ надо уметь

защищать не только от «зарубежных партнеров», но и от коллег из собственной страны.

Следует сказать, что в области ответственных приложений часть работ велась отдельно от отечественных программ по освоению СКТ. Создание адекватной требованиям таких приложений зарубежной техники постоянно отслеживалось все эти годы [23,24,25], а в качестве ответных мер достаточно успешно использовалось применение FPGA и разработка специализированной ЭКБ в виде проблемно-ориентированных СБИС [26,27,34,20]. Полученный опыт и наработки могут быть использованы при постановке работ по созданию ЭКБ ИИ.

2. Новые вызовы ЭКБ ИИ и их значимость

На сегодняшний день популярным изделием на отечественном рынке являются сервера NVIDIA DGX2, содержащие по 16 GPU V100, соединенных через специальные коммутаторы линков NVlink. Они имеют пиковую производительность на FP16/FP32 около 2 PF, также содержат 2 скалярных процессора Intel Xeon Platinum для управления, 1 TB оперативной системной памяти и энергонезависимую память в 30 TB на приборах с зарядовой связью, мощные интерфейсы Mellanox для связи с другими такими серверами.

Такие сервера служат в настоящее время основным строительным блоком суперкомпьютеров для построения систем ИИ кластерного типа. Однако стоит такой сервер около $400 тыс. (₽27 млн), а для реальных приложений их требуется много. Например, стоимость лишь только таких серверов суперкомпьютера кластерного типа Сбербанка оцениваются не менее, чем в ₽2 млрд. Заметим для сравнения, что разработка и изготовление инженерных образцов Эльбрус 16С в трехлетнем проекте стоило ₽1,5 млрд. Может быть, на отечественный GPU тем же разработчикам или их коллегам не надо было жалеть денег? Наверняка в ближайшее время похожих кластеров для ИИ потребуются десятки, но про собственное ЭКБ для ИИ пока не говорят. Вот такое в итоге «импортозамещение», это иллюстрация последствий допущенных ошибок освоения технологий СКТ.

Тем не менее, рынок ЭКБ даже уже сейчас разнообразен и динамично меняется. Если учитывать только средства работы с нейронными сетями, причем глубокого обучения (DNN), то на нем присутствуют не только GPU V100 (применяются, в основном, для обучения нейросетей), но и процессоры Intel, они используются для вывода решений с применением уже обученных нейросетей.

В работе [21] даются приведенные на рис. 1 сведения по состоянию рынка на 2017 г. и прогноз на 2025 г. Видно, что выделены два сегмента – аппаратные средства центров обработки данных (data-center architecture, %, ЦОД-ы) и автономные средства (edge architecture, %). В каждом из этих

сегментов отдельно рассмотрены части средств для обучения нейросетей (training) и решения задач на обученных сетях (inference).

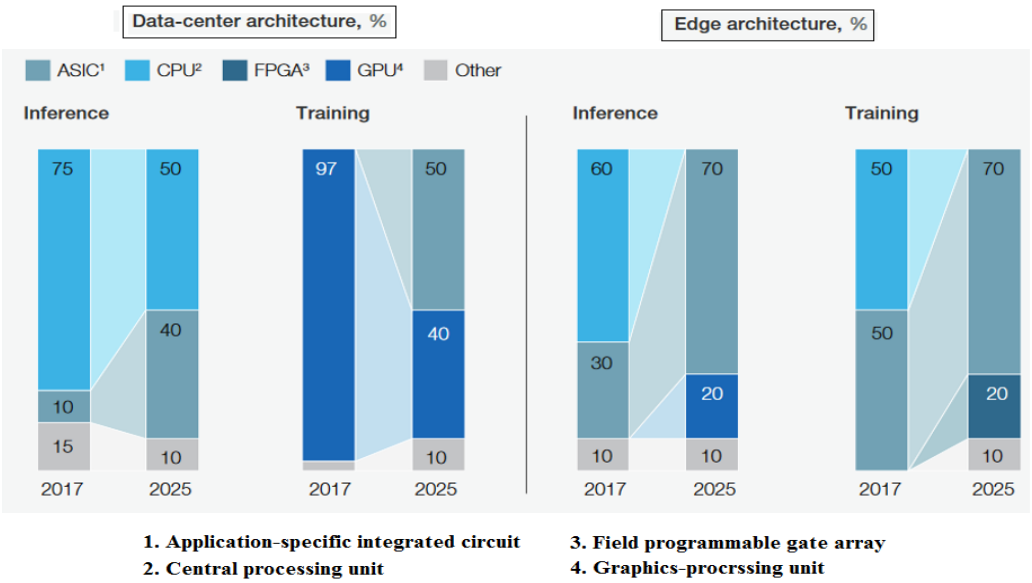


Рис. 1. Качественный состав и динамика изменения рынка ЭКГ сетей типа DNN

Представление об объемах ранка выделенных сегментов и их частей дает рис. 2, а вот общая значимость обоих сегментов по их доле в мировом рынке полупроводников приведена на рис. 3. Впечатляет оценка рынка связанной с ИИ полупроводниковой продукции, в 2025 г. – это $65 млрд. Для сравнения, мировой рынок вооружений в 2019 г. составлял $92 млрд.

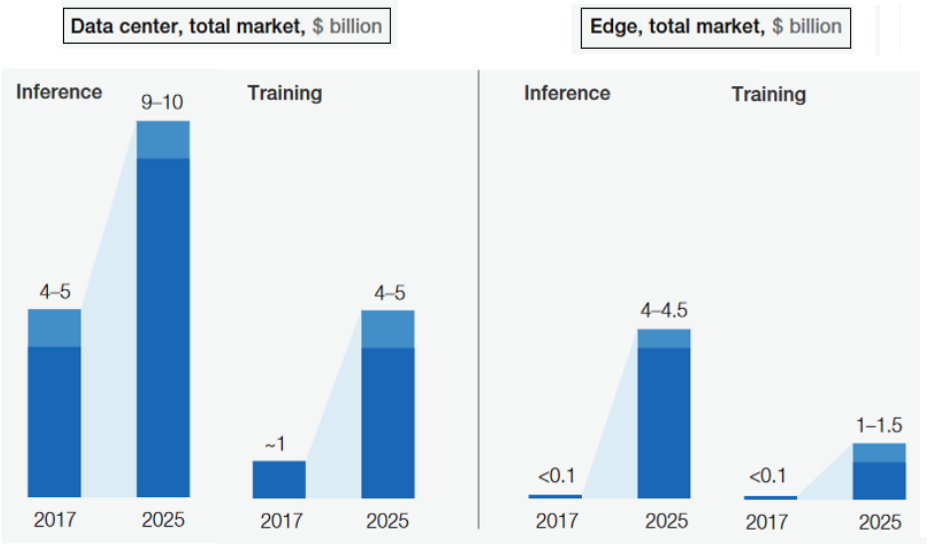


Рис. 2. Объем рынка только аппаратных средств работы с нейросетями типа DNN

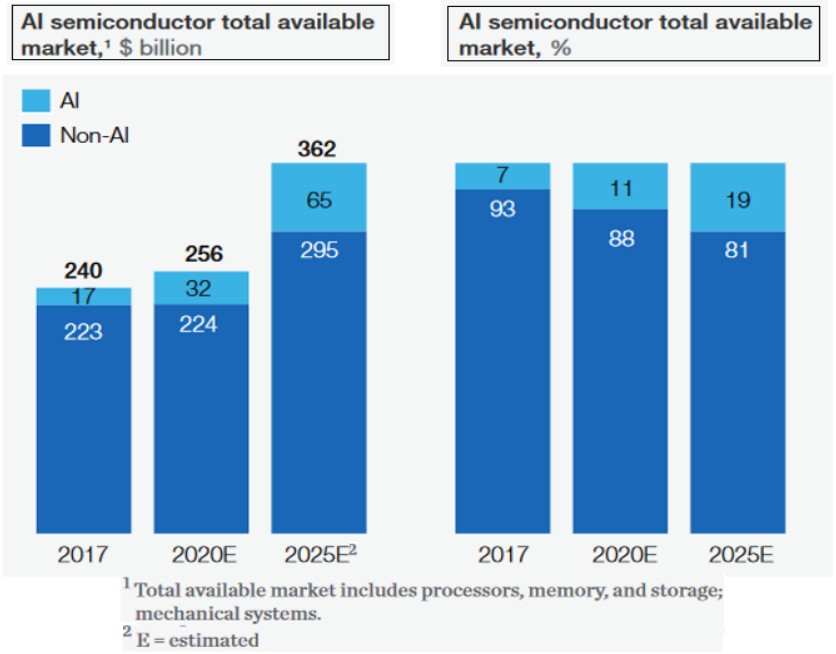


Рис. 3. Объемы мирового рынка полупроводниковой продукции и ее части, связанной с изделиями с элементами искусственного интеллекта

Далее вернемся к рис. 1 с целью рассмотрения, какие типы процессоров для обучения и вывода используются сейчас и что ожидается в перспективе в 2025 г. При этом будем учитывать значимость таких изменений по данным рис. 2. Можно сделать следующие выводы.

1. В аппаратных средствах для тренировки и вывода одного и другого сегмента значительная доля (от 40% до 70%) приходится на специализированные процессоры (ASIC). Наиболее заметно это произойдет на рынке автономных средств (70%). Более впечатляет прогноз, что ASIC-процессоры потеснят GPU, абсолютного лидера на сегодняшний день в системах обучения сегмента ЦОД. На долю GPU приходится сейчас 97%, а прогнозируется снижение до 40%, при этом доля ASIC-процессоров должна подняться до 50%.

Конкуренцию GPU и ASIC в текущее время можно увидеть по происходящим сравнениям характеристик как отдельных кристаллов GPU V100 и Gaudi, так и многосокетных блоков на их основе NVIDIA DGX-2 и HLS-1 (8 Gaudi), а также кластерных суперкомпьютеров на их основе [22]. Надо сказать, что результаты сравнения пока не в пользу NVIDIA, особенно по масштабированию производительности. Для кластера на GPU V100 она резко падает после 16 GPU и на 512 GPU составляет немного больше 20% от пропорционально увеличенной пиковой. Для кластера на процессорах Gaudi на 512 процессорах производительность 80%.

Вместе с тем, все еще впереди, как указано в [22], фирма NVIDIA имеет ряд преимуществ на перспективу, это выпуск нового GPU Ampere, использование более мощных линков от Mellanox (эту израильскую фирму

NVIDIA купила за $6,9 млрд), накопленное мощное программное обеспечение, возможность более разнообразного использования в силу большей универсальности GPU в сравнении с ASIC.

2. Несмотря на прогнозируемый проигрыш GPU в конкурентной борьбе ASIC-процессорам, GPU будут иметь 40% в сегменте ЦОД (обучение) и 20% в сегменте автономных устройств (вывод). Очевидно, что в силу своей универсальности GPU по-прежнему будут использоваться в научно-технических расчетах, да и еще в таких областях, которые сейчас даже трудно предсказать. В связи с этим вывод – работы по разработке отечественного варианта GPU надо продолжать и в рамках работ по ЭКБ ИИ, используя все возможные варианты [20].

3. Видно, что перспектив использования CPU в автономных устройствах нет ни при обучении, ни при выводе. Однако в сегменте ЦОД (вывод), а это значительная доля рынка, доля CPU предсказывается в 50%. Это оценка роли обычных скалярных процессоров в данной области, для нашей страны – это место Эльбрус 16С и последующих образцов этого семейства.

4. Программируемые интегральные схемы (FPGA), которые в настоящее время в нашей стране особенно популярны, могут занять в будущем лишь 20% очень скромного рынка средств обучения нейросетей в сегменте автономных устройств. Это следует учесть, поскольку в настоящее время FPGA рассматриваются как решение всех проблем с ЭКБ в любых системах.

Общий вывод – сильные перспективы имеет разработка ASIC-процессоров для одного и другого сегмента, как при обучении, так и при выводе. На втором месте GPU. Это необходимо учитывать при разработке программы по освоению технологий ИИ в части ЭКБ ИИ. Как на это повлияют грядущие изменения в мировой экономике, пока непонятно. Скорее всего, мировая значимость рассматриваемой области ЭКБ для ИИ возрастет на фоне ослабления традиционных секторов экономики, особенно сырьевой.

Если детализировать тему ASIC-процессоров в качестве ЭКБ ИИ, то можно выделить следующие направления:

– процессоры для работы с DNN сетями, здесь основная задача состоит в реализации высокопроизводительных и энергоэффективных структур арифметических устройств умножения плотнозаполненных матриц, результаты достигнуты уже фантастические, для этого применяются плоские большие систолические матрицы арифметических устройств (TPU Google, Goya, Gaudi [22,28]), трехмерные структуры арифметических устройств (Nervana NNT-I, Huawei DaVinchi), применяются разные подходы подкачки данных, включая скалярно-векторные и мультитредовые архитектуры в отдельных ядрах, это известный принцип

DAE (decoupled access/execute architecture, разделения работы с данными и вычислений);

– процессоры для работы с SNN сетями, это непосредственные модели нейронных сетей, с представлениями ядер нейронов, аксонов, дендритов и синапсов, моделирования во времени распространения между ними спайков (импульсов), пока очень многообещающее и сложное направление (TrueNoth [29], Loihi [30]), которое в сравнении с DNN сетями может дать лучшее качество, а главное – на 3 порядка лучшую энергоэффективность, поскольку таких вычислений, как в DNN, не требуется, но у человека $89 \cdot 10^9$ нейронов и 10^{14}–10^{15} синапсов (точек соединения аксонов нейронов с дендритами, в которых должно быть обучение), как это все поместить в кремнии и суперкомпьютерах и надо ли все, это область активной работы с нейрофизиологами;

– процессоры одновременно с сетями DNN и SNN, известен пример такого процессора Tianjin университета в Пекине для управления велосипедом;

– процессоры с DNN и/или SNN и с множеством специальных блоков и процессорных ядер для управления роботами и транспортными средствами, пример – NVIDIA Xavier, DARPA HIVE;

– - массово многоядерные процессоры с мощной внутрикристальной сетью, мультитредовыми ядрами, включающими локальную память и специализированные блоки-ускорители вычислений [31,32], это очень перспективное направление, которое позволит одновременно реализовать DNN и SNN сети на единой аппаратной основе, также в обобщенном виде этот вариант применим и как вариант отечественной реализации противопоставляемого NVIDIA GPU отечественного микропроцессора (выделен как 3-й вариант решений по направлению 1 в [6]), по этому варианту уже ведется инициативная разработка в виде открытого учебного проекта [20].

Процессоры типа GPU и ASIC содержат сложно-функциональные блоки (IP-блоки), которые либо характерны именно для ЭКБ ИИ, либо обычно применяемые, но с форсированными характеристиками. Такие IP-блоки применимы при разработке разных образцов GPU и ASIC и могли бы разрабатываться отдельно в большей среде не столь многочисленных коллективов, чем коллективы-разработчики GPU и ASIC. Разработка похожих IP-блоков актуальна и для технологий СКТ, что отмечено в [6] как направление 4.

Можно выделить для начала следующие направления разработки IP-блоков и связанных с ними технологий:

– функциональные устройства выполнения операций над плотнозаполненными матрицами, это сейчас основной элемент выпускаемых микропроцессоров, имеется большое количество патентов и статей, надо экспериментировать, а существующие решения впечатляют,

это систолические матрицы (систолические вычислительные системы – системы, основным принципом которых является то, что все данные регулярно и ритмично проходящие через массив, используются многократно) размера 256×256, 128×128, трехмерные структуры 4×4×4, 16×16×16, а уровень параллелизма, который они обеспечивают, – несколько тысяч операций за такт;

– функциональные устройства для выполнения операций над разреженными матрицами, это очень актуальная тема исследований и разработок, такие устройства реализуются в проектах процессоров для работы с графами нового поколения [3,4,5], исследуются с целью дальнейшего применения в нейропроцессорах [33], важны также для нового поколения вычислительных средств выполнения научно-технических расчетов, например, на нерегулярных динамически изменяемых сетках;

– внутрикристальные сети с топологией «решетка», «2D-тор» и в перспективе с топологией сети с ограниченным диаметром (направление 3 в работе [6]);

– технологии 3D-сборки для изготовления HBM-модулей памяти с повышенной пропускной способностью, а также PIM-технологии (процессоры в памяти) гибридных 3D сборок с чипами HBM-памяти и процессоров;

– технологии высокоскоростных и энергоэффективных SER/DES приемопередатчиков для реализации межкристальных линков.

До сих пор в статье рассматривалась лишь тема ЭКБ, поскольку она наиболее пострадала в похожем на затеваемый процесс освоения технологий ИИ процесс освоения СКТ. Однако в данном случае беспокойство вызывает и тема алгоритмов в области ИИ. Дело в том, что при освоении СКТ был резерв специалистов в области физики и математики, воспитанных в советский период в знаменитых школах с мировым именем. Что же касается алгоритмов ИИ, то в отечественной экспертной среде есть опасения, что в нашей стране многое упущено, тем более, что в данном случае необходимо подключение и нейрофизиологов. Выйти на зарубежный уровень будет крайне сложно, если вообще возможно. Будем надеяться, что сохранившийся еще научный потенциал при правильной организации работ по ИИ решит и эту проблему.

В заключение еще сделаем замечание о возможной новой организации работ предприятий-разработчиков процессорных чипов, которая связана именно с отраслью ЭКБ для ИИ. Идея такой новой организации была изложена в [21] и показалась авторам интересной. Она состоит в следующем.

Авторы работы [21] отмечают, что выпуск в предыдущие годы персональных компьютеров и мобильных телефонов принес полупроводниковым фирмам-изготовителям чипов всего лишь 10% от

прибыли. Основную долю получили разработчики программного обеспечения и, собственно, продаваемых изделий. Напомним, что это было время, когда улучшение характеристик изделий шло за счет совершенствования технологий, исследования по архитектуре и микроархитектуре были не так нужны.

В предстоящий период разработка чипов потребует больших усилий именно в части их архитектуры и микроархитектуры, поскольку усовершенствование КМОП-технологий дошло до предела. Тем не менее, предполагаемый масштаб и разнообразие изделий на рынке ЭКБ ИИ может позволить получать именно фирмам-изготовителям чипов до 40-50% прибыли, но для этого им надо взять на себя многие функции изготовителей конечного оборудования, системного и прикладного программного обеспечения. Для этого предлагается организовать вертикальные структуры, в основе которых и будут фирмы-изготовители чипов. Собственно говоря, это логично из-за ожидаемой быстрой сменяемости продукции на рынке ИИ и уже происходит, что заметно в настоящее время.

Заключение

В статье были рассмотрены темы для обсуждения концепции ФЦП и ЦПСГ по освоению и развитию технологий ИИ. Особое внимание было уделено, допущенным, по мнению авторов, ошибкам в длившемся последние 20 лет процессе освоения и развития СКТ. Имеется ввиду тема запущенности при этом развития ЭКБ. Для ЭКБ ИИ такая ситуация абсолютно недопустима, поскольку для рынка ИИ роль ЭКБ ожидается доминирующей.

В силу стратегической важности темы, предлагается обсудить еще следующие три предложения, касающиеся организации таких работ.

Во-первых, предлагается взять элементы организации атомного проекта 1940-50-х гг. Можно было бы сформировать Дирекцию программы (далее – Дирекция) и назначить персонально ответственного за нее перед правительством, установить регулярную отчетность. При Дирекции организовать Научный совет. Осуществлять конкретное планирование работ и контроль их выполнения с привлечением членов Научного совета. Организовать при Дирекции мощную информационно-аналитическую службу, которая бы работала в тесном взаимодействии с Научным советом и информационно-аналитическими службами исполнителей программы. Сразу же предусмотреть меры по противодействию внешним и внутренним деструктивным силам, что может выполняться отдельной службой внутри Дирекции, которая была бы в тесном контакте с назначенными управлениями федеральных служб.

Во-вторых, цели программы, решаемые задачи для достижения этих целей, планы работ должны быть конкретно прописаны, а не выбираться

так, как принято в последнее время по принципу «кто что может сделать» для решения достаточно в общем виде описанной проблемы. По-видимому, можно установить два направления работ:

– построение, изучение, использование кластерных суперкомпьютеров для решения задач ИИ на базе зарубежных технологий;

– разработка собственной ЭКБ для систем ИИ и заодно для СКТ в части решения проблемы импортозамещения, а более конкретно, это отечественный GPU и ASIC, варианты решений, например, как в работе [20], выполнение работ по необходимым IP-блокам.

В-третьих, исполнителей работ выбирать не по результатам тендера с подсчетами баллов, а с учетом реальных возможностей организаций, наличию школ и опытных коллективов, индивидуального выбора лидеров работ. Это должна выполнять изначально сформированная Дирекция и Научный совет программы.

Реализация программы могла бы начаться с некоторых пробных, менее масштабных проектов. Наиболее актуально сейчас – разработка отечественного GPU [20], начать можно с ФЦП, потом расширить фронт работ.

Литература

1. *Эйсымонт Л.К.* Компьютеры для обработки символьной информации // Зарубежная радиоэлектроника. 1990, №4, с.3-28.
2. *Эйсымонт Л.К.* О возможности параллельных схем реализации одного языка для описания задач переработки текстовой информации // Управляющие Системы и Машины, Киев, 1977. С.56-64.
3. *Song W.S., Gleyzer V., Lomakin A., Kepner J.* Novel Graph Processor Architecture, Prototype System, and Results // 2016 IEEE High Performance Extreame Computing Conference, 22 July 2016. http://arhiv.org/abs/1607.06541
4. *Song W.S.* Processor for large graph algorithm computations and matrix operations // US Patent No 9,529,590 B2, Dec 27, 2016.
5. *Dai G. et al*. GraphH: A Processing-in-Memory Architecture for Large-scale Graph Processing // IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, April 2019, Vol.38, N4, p.640-653. https://cseweb.ucsd.edu/~jzhao/files/GraphH-tcad.pdf
6. *Адамов А.А., Фомин Д.В., Эйсымонт Л.К.* Главные проблемные направления в области отечественной элементной базы суперкомпьютеров // Вопросы кибербезопасности. 2019, №4, с.2-12.
7. *Задыхайло И.Б., Камынин С.С., Любимский Э.З.* Вопросы конструирования вычислительных машин из блоков повышенной квалификации. – Препринт ИПМ им. М.В. Келдыша АН СССР. 1971, №68.

8. *Митрофанов В.В., Слуцкин А.И., Ларионов К.А., Эйсымонт Л.К.* Вычислительные кластеры и проблемы развития отечественных высокопроизводительных систем // Системы и средства связи, телевидения и радиовещания. 2001, вып.2-3, с.30-35.
9. *Митрофанов В.В., Слуцкин А.И., Ларионов К.А., Эйсымонт Л.К.* Направления развития отечественных высокопроизводительных систем // Открытые системы. 2003, №5, с.29-35.
10. *Митрофанов В.В., Эйсымонт Л.К.* Элементная база и архитектура высокопроизводительных мультипроцессрных вычислительных систем, перспективных стратегических и встроенных суперкомпьютеров // Динамика радиоэлектроники, 2 выпуск. – М.: Техносфера, 2008. С.70-76.
11. *Konecny P.* Introducing Cray XMT // Proc. Cray User Group meeting (CUG 2007), May 2007 – CUG Proceedings.
12. *Слуцкин А.И., Эйсымонт Л.К.* Российский суперкомпьютер с глобально адресуемой памятью // Открытые системы. 2007, №9, с.42-51.
13. *Митрофанов В.В., Слуцкин А.И., Эйсымонт Л.К.* Суперкомпьютерные технологии для стратегически важных задач // Электроника: НТБ. 2008, №7, с.66-79.
14. *Семенов А.С., Соколов А.А. Эйсымонт Л.К.* Архитектура глобально адресуемой памяти мультитредово-потокового суперкомпьютера // Электроника: НТБ. 2009, №1, с.50-56.
15. Кудрявцев М., Эйсымонт Л., Мошкин Д., Полунин М. Суперкластеры – между прошлым и будущим // Открытые системы. 2008, №8. https://www.osp.ru/os/2008/08/5661383/
16. *Речинский А., Горбунов В., Эйсымонт Л.* Суперкластер с глобально адресуемой памятью // Открытые системы. 2011, №7, с. 21-25.
17. Whitepaper NVIDIA’s Next Generation CUDA Compute Architecture: Fermi – 2009 – NVIDIA Corporation.
18. *Фролов А.С., Семенов А.С., Корж А.А., Эйсымонт Л.К.* Программа создания перспективных суперкомпьютеров // Открытые системы. 2007, №9, с.20-29.
19. *Dongarra J. et al.* DARPA's HPCS Program: History, Models, Tools, Languages. http://users.sdsc.edu/~lcarring/Papers/2008_AC.pdf
20. *Адамов А.А., Павлухин П.В., Биконов Д.В., Эйсымонт А.Л., Эйсымонт Л.К.* Альтернативные современным GPGPU перспективные универсальные и специализированные процессоры-ускорители // Вопросы кибербезопасности. 2019, №4, с.13-21.
21. *Batra G., Jacobson Z., Madhav S., Queiro A., Santham N.* Artificial intelligence hardware: new opportunities for semiconductor companies // McKinsey&Company, Dec 2018.

22. *Gwennap L.* Habana offers Gaudi for AI Training. Startup Expects to top Nvidia V100 Performance at Half the Power // Microprocessor Report, June 2019.
23. *Горбунов В., Эйсымонт Л.* Экзафлопсный барьер: проблемы и решения // Открытые системы. 2010, №6, с.12-15.
24. *Горбунов В., Елизаров Г., Эйсымонт Л.* HPC: региональные новости. // Открытые системы. 2011, №2, с.12-16.
25. *Горбунов В.С., Елизаров Г.С., Эйсымонт Л.К.* Экзафлопсные суперкомпьютеры: достижения и перспективы // Открытые системы. 2013, №7, с.10-14.
26. *Эйсымонт Л.К.* Гибридная стратегия развития элементной базы // Открытые системы. 2017, №2. https://www.osp.ru/os/2017/02/13052216/
27. *Эйсымонт Л.К.* Настраиваемые специализированные СБИС – реальная основа создания будущих экзамасштабных суперкомпьютеров, зарубежный и отечественный опыт // Системы высокой доступности. 2018, т.14, №3, с.18-27.
28. Gaudi Training Platform White Paper // Habana Labs, June 2019.
29. *Akopyan F. et al*. TrueNoth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip // IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. October 2015, Vol.34, N10, pp.1537-1557
30. *Davies M., Srinivasa N., Lin T.H. et al.* Loihi: A neuromorphic manycore processor with on-chip learning // IEEE Micro 38 (1), p.82-99
31. *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.
32. PEZY-SC2-PEZY// https://en.wikichip.org/wiki/pezy/pezy-scx/pezy-sc2
33. *Han S., Liu X., Mao H., Jing Pu.J., Pedram A., Horowitz M.A., Dally W.J.* EIE: Efficient Inference Engine on Compressed Deep Neural Network // 2016 ACM/IEEE 43rd Annual ISCA.
34. *Елизаров С.Г. и др.* Программируемые на языках высокого уровня энергоэффективные специализированные СБИС для решения задач информационной безопасности // Системы высокой доступности. 2018, т.14, №3, с.40-48.