НИЦЭВТ

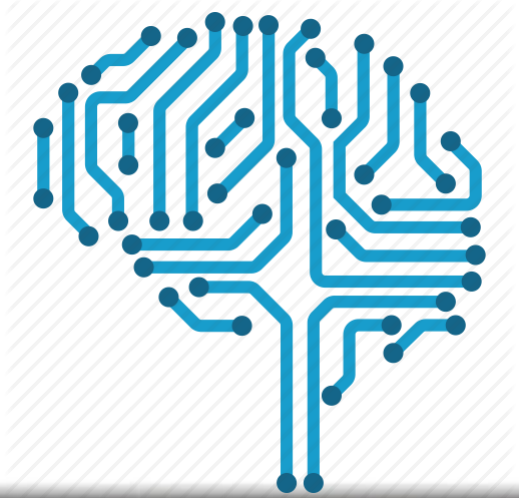
Ростех

Тенденции развития технологий Big Data и высокоскоростная сеть Ангара

А.С. Семенов, А.С. Фролов

1.03.2018

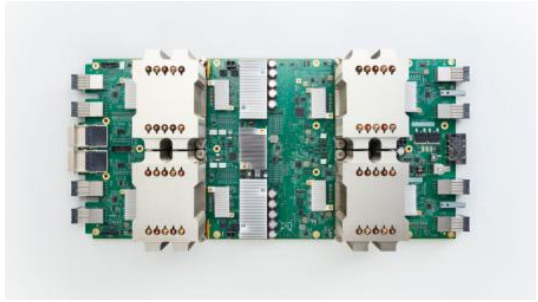
Современный мир



Вычислительные технологии в нейронных сетях

Обучение

Применение



Google TPUv2

NVIDIA Volta

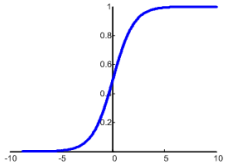
Google TPU

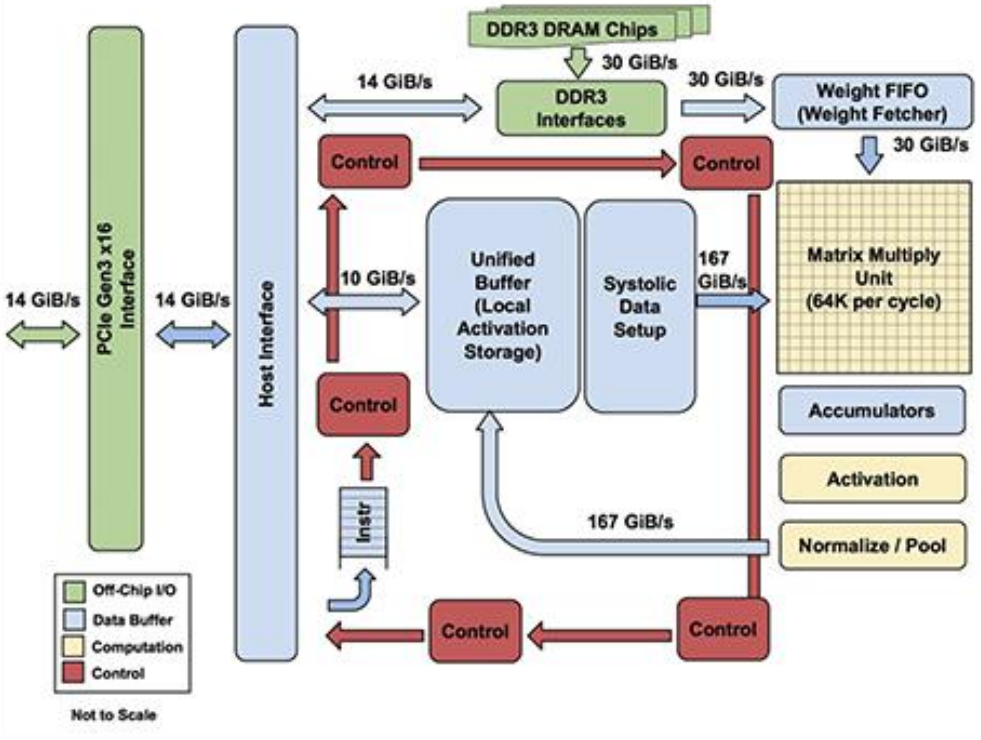
Функциональность vs Производительность

Google Tensor Processing Unit (~2013 – 2015)

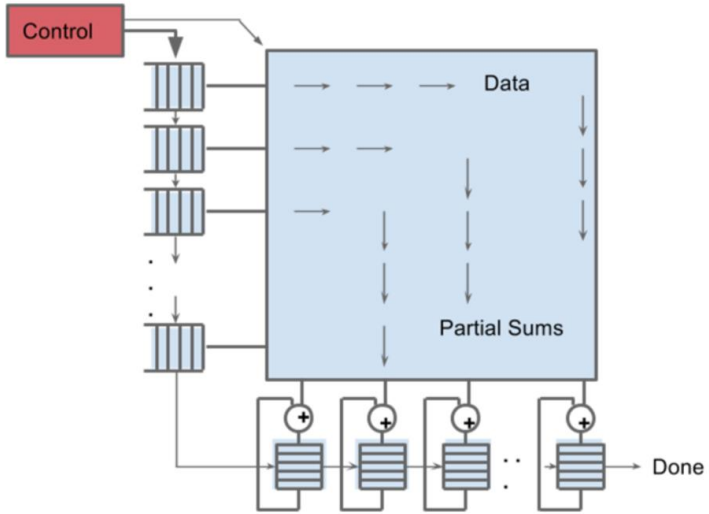
- Matrix Multiply Unit (MXU) – 64K 8 bits fma
- Activation Unit – функции активации
- ~ 12 инструкций CISC
  - Read_Host_Memory
  - Read_Weights
  - MatrixMultiply/Convolve
  - Activate
  - Write_Host_Memory

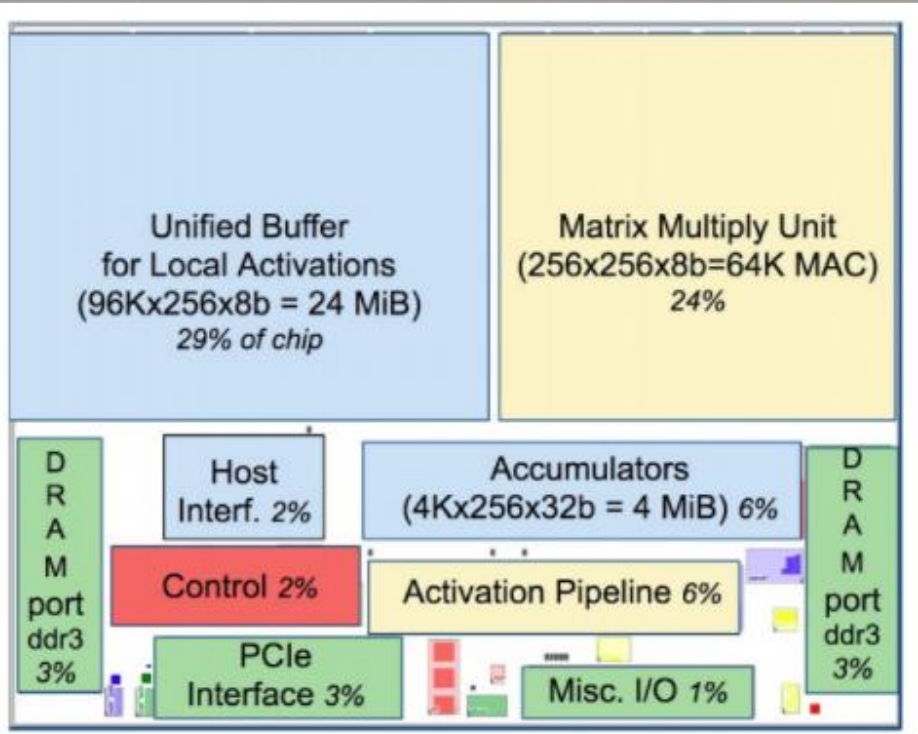
$$f(x) = \frac{1}{1+e^{-x}}$$





Google Tensor Processing Unit





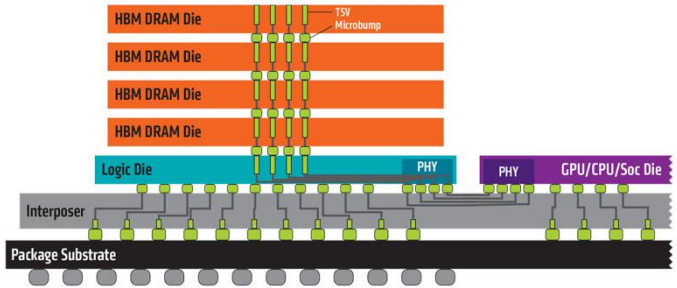
- 28 nm, 700 MHz, 92 Tops, 40W
- **Производительность/W**:
  83X vs CPU, 29X vs GPU
- Unified Buffer – 24 MB
- 8 GB DDR3, 34 GB/s

Google Tensor Processing Unit v2 (2017)

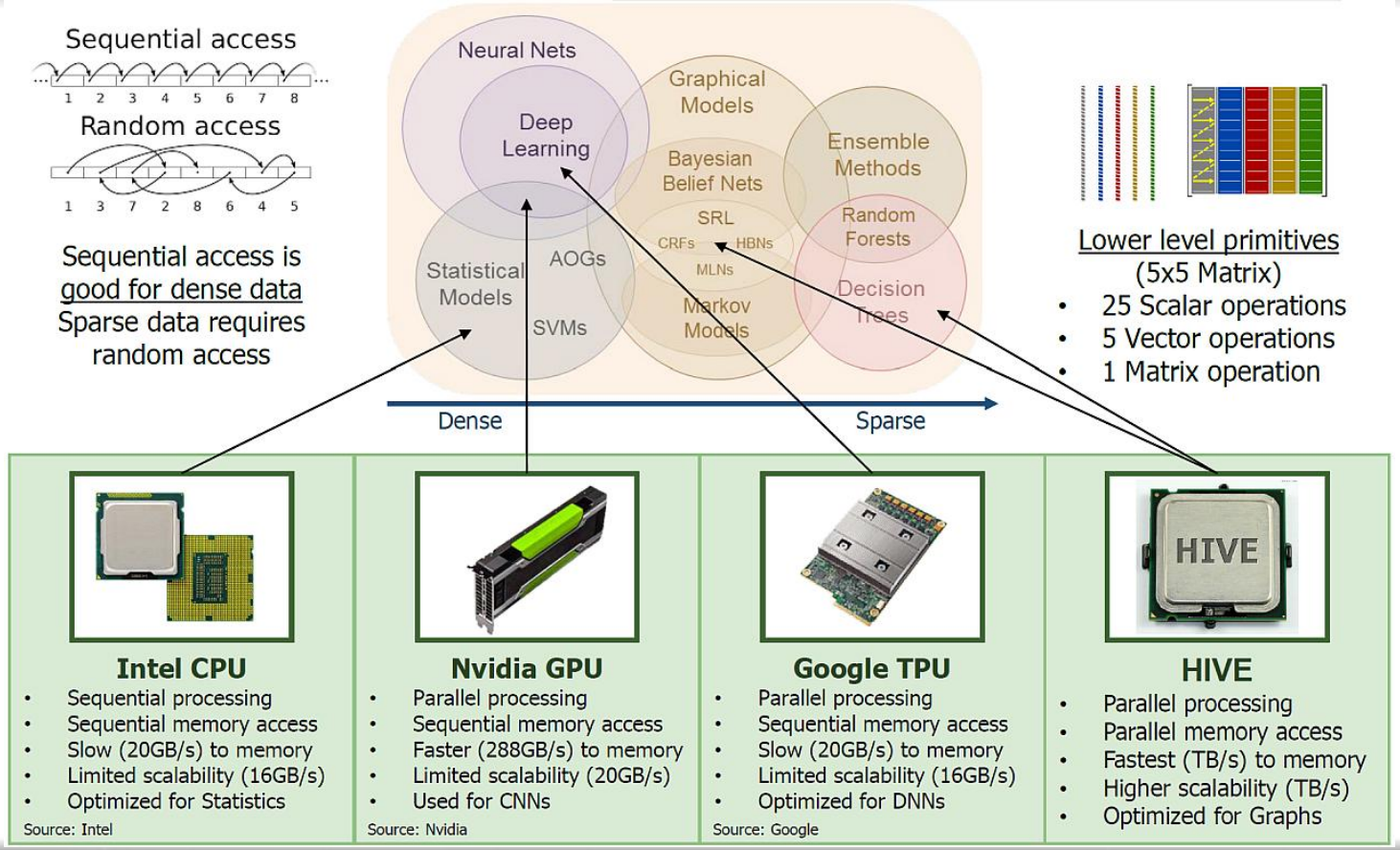
- TPU v1 – **bandwidth limited**

TPU v2

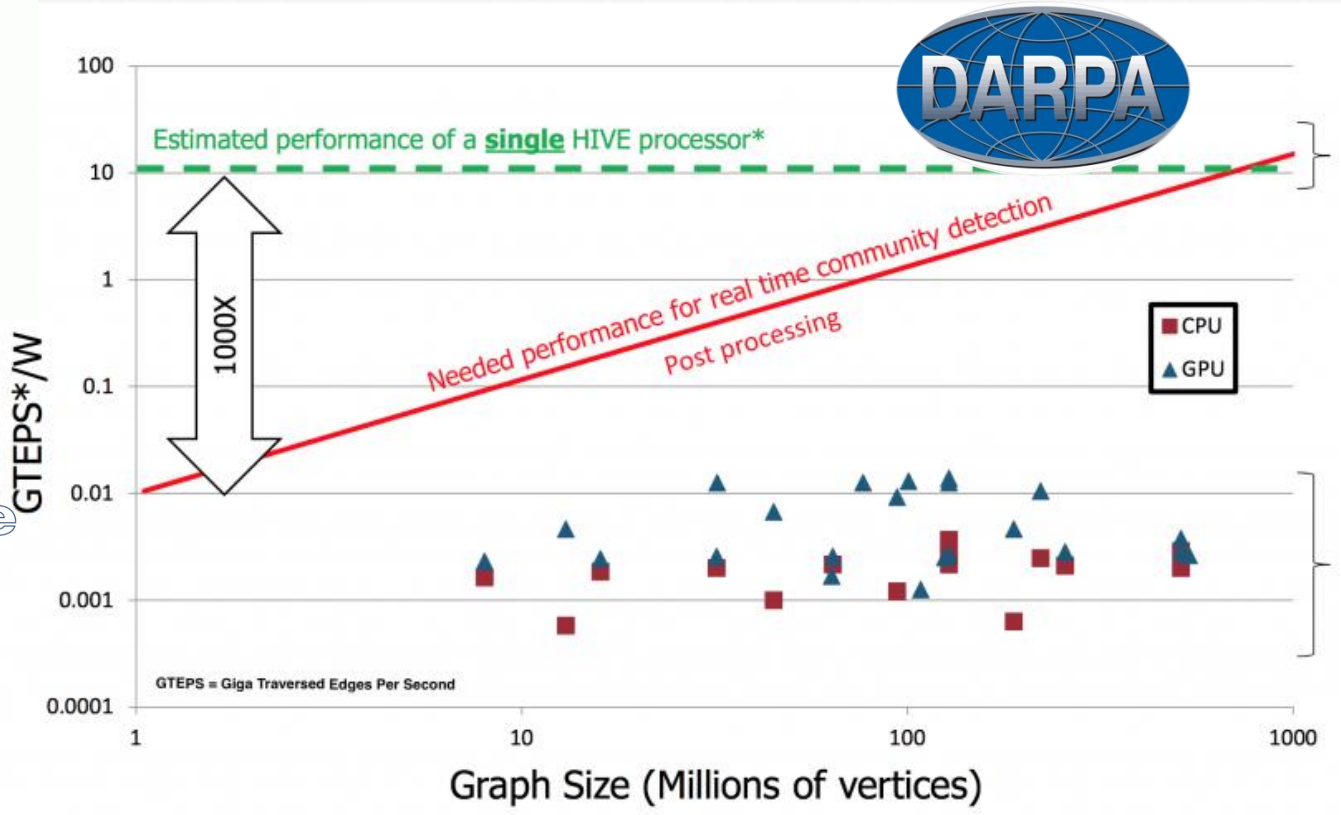
- HBM 16 GB – 600 GB/s
- 32 bits floating point
- Обучение и применение
- 45 TFlops



Вычислительные элементы
Нейронные сети vs Графы



DARPA HIVE: графовый процессор (2016-2020)



Архитектурные особенности процессора обработки графов

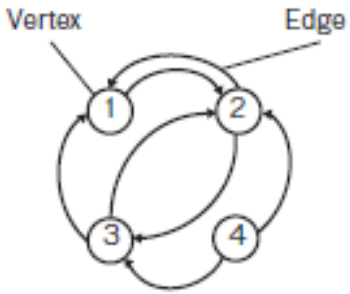
НИЦЭВТ

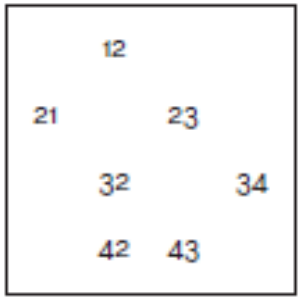
Ростех

GraphBLAS

MIT Lincoln Laboratory

Операция	Описание
A*x, A*B	Обход графа (BFS), поиск кратчайших путей
A+B, …	Объединение, пересечение графов



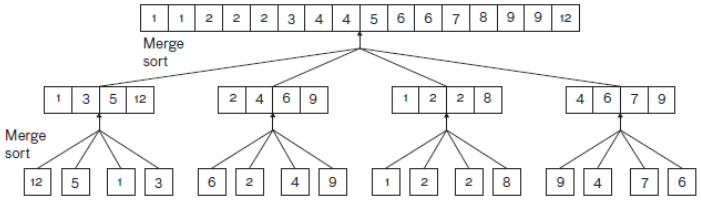


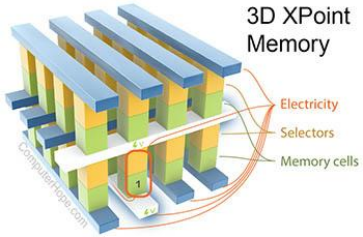
Coordinate format storage (row based)

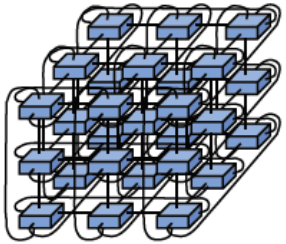
Data value	12	21	23	32	34	42	43
Row index	1	2	2	3	3	4	4
Column index	2	1	3	2	4	2	3

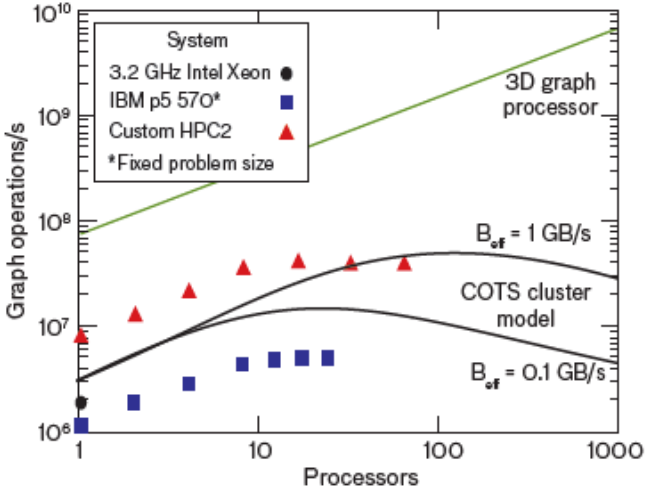
95% операций – сортировка

k-way merge sort

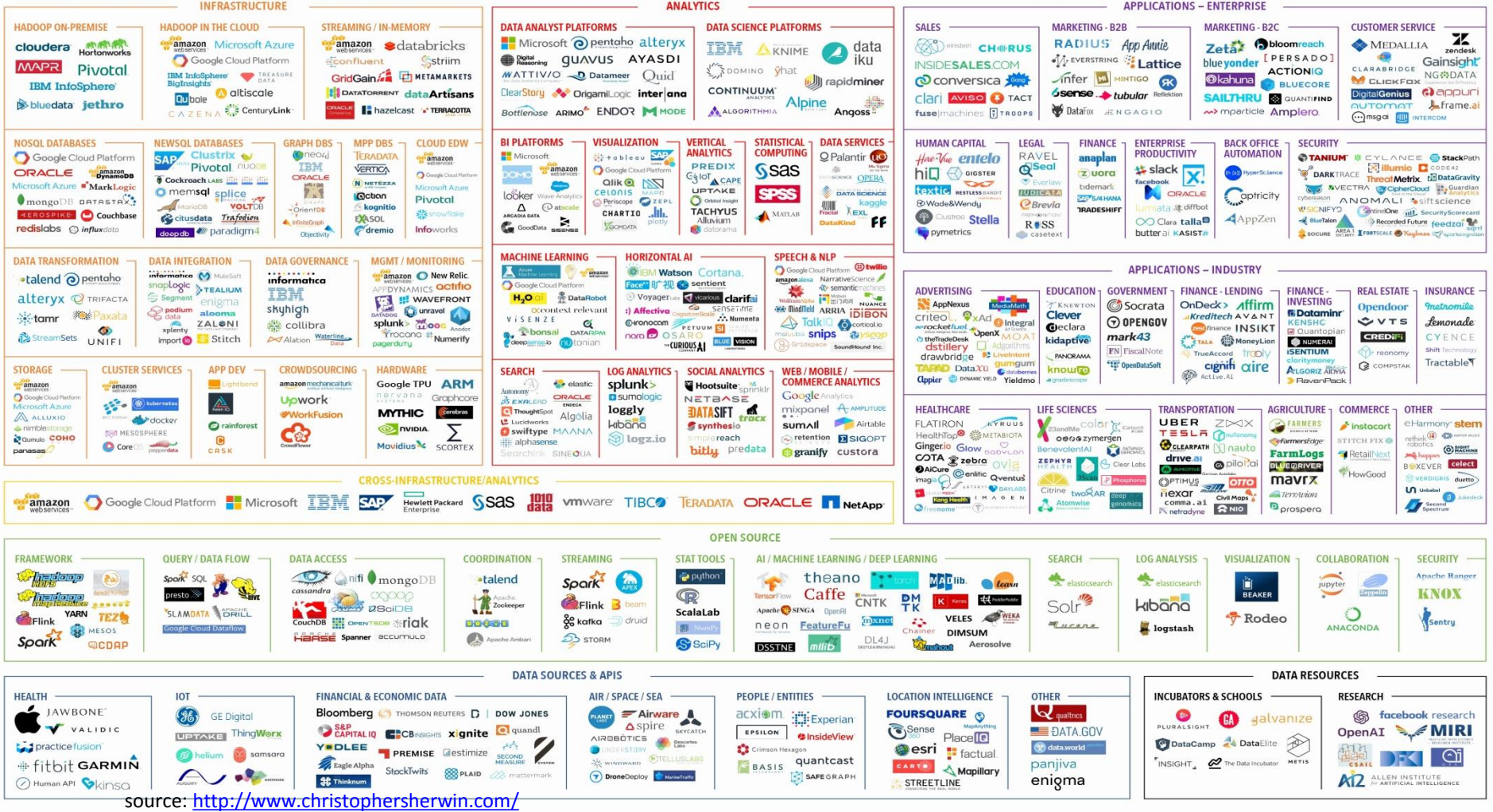








BIG DATA LANDSCAPE 2017



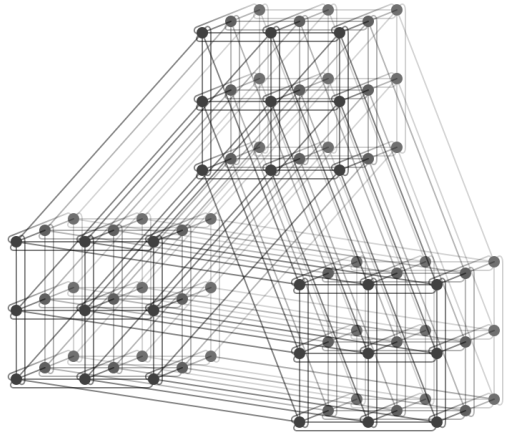
source: http://www.christophersherwin.com/

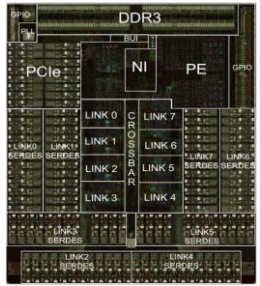
HPC vs BigData

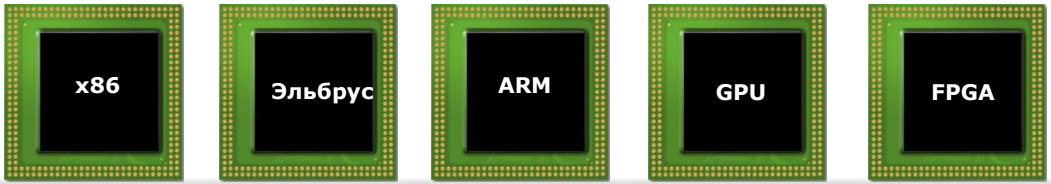
	HPC	BigData
Языки программирования	C/C++, перекомпиляция	JVM, Python, переносимость
Типы нагрузки	Изначально вычисления	Много и сложных данных
Масштабируемость	Изначально вертикальная	Горизонтальная
Управление выполнением	На каждом узле – MPI	Раздача заданий – Spark jobs
Файловые системы	Lustre, …	HDFS, локальные данные
Аппаратура	High-end, RDMA	Commodity, Ethernet
Отказоустойчивость	Необходима обработка	Встроенная

Ключевые особенности:

- Топология сети: 1D..4D-тор
- Адаптер на базе СБИС (65 нм, АО «НИЦЭВТ»)
- До 8 каналов связи с соседними узлами
- Прямой доступ в память удаленного узла (RDMA)
- Поддержка многоядерности
- Адаптивная передача пакетов
- Задержка на MPI ping-pong: 0,85/ 1,54 мкс (x86/Эльбрус-8С)
- Задержка на хоп: 130 нс
- Масштабирование: до 32K узлов
- Энергопотребление до 20 Вт
- Различные физические среды передачи данных





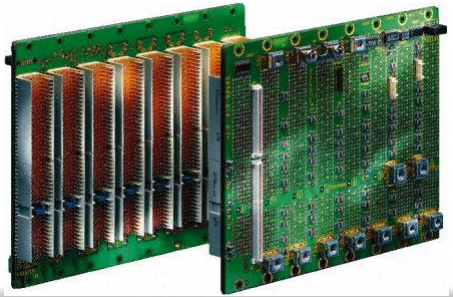


Варианты сетевого оборудования Ангара

1. Высокопроизводительное решение на базе FHFL адаптера и Samtec кабеля

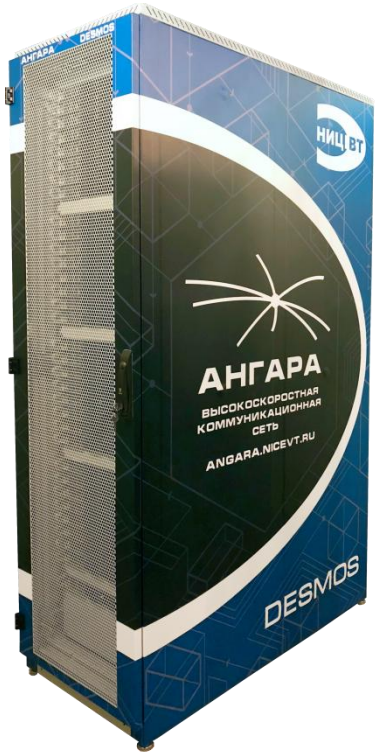
2. Универсальное решение на базе 24-портового коммутатора, low-profile адаптера и CXP кабеля

3. Заказное решение на базе объединительной платы и оптических кабелей

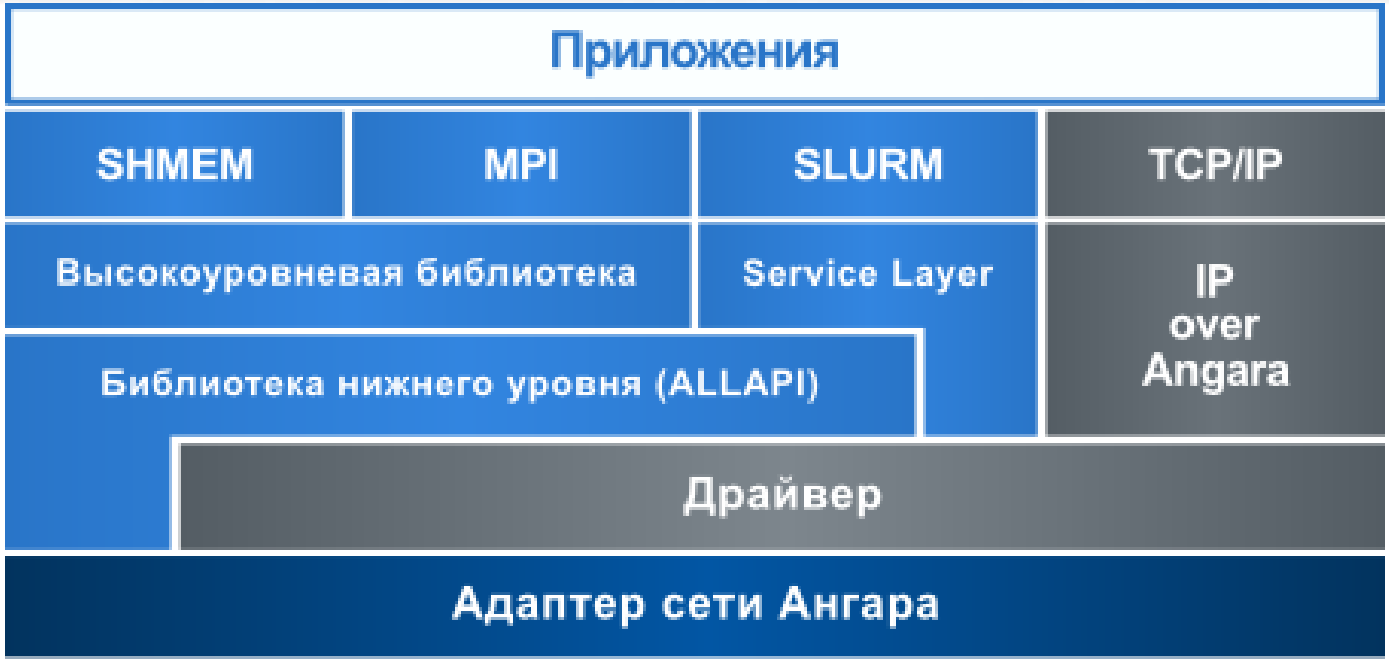


Инсталляции

- **ОИВТ РАН: 32 вычислительных узла**
  - 1 процессор Intel Xeon E5-1650 v3 (6 ядер, 3.0 ГГц)
  - Nvidia GeForce GTX 1070
  - DDR4 16 ГБ
  - 4D-тор 4x2x2x2
- **Ангара-К1: 36 вычислительных узлов**
  - 12 узлов с 1 процессором Intel Xeon E5-2660 (8 ядер, 2.2 ГГц)
  - 24 узла с 2 процессорами Xeon E5-2630 (6 ядер, 2.3 ГГц)
  - 64 ГБ
  - 3D-тор 4x3x3



Стек программного обеспечения



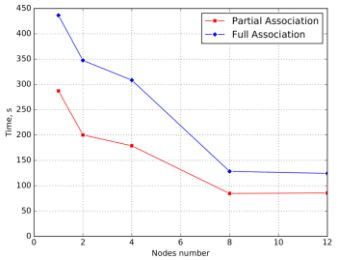
- Поддержка ОС: Astra Linux SE 1.3-1.5, ОС «Эльбрус», OpenSUSE/SLES 11 SP3/4, CentOS 6.0-7.3, Версия ядра Linux от 2.6.21 до 3.16.0
- MPICH 3.0.4, 3.2, OpenMPI 1.10.2
- Поддержка компиляторов языков Fortran 77/90/95 (GNU, Intel), C/C++ (GNU, Intel)

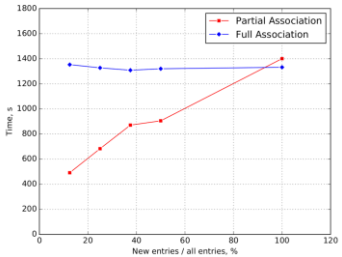
ANSYS FlowVision OpenFOAM The Open Source CFD Toolbox

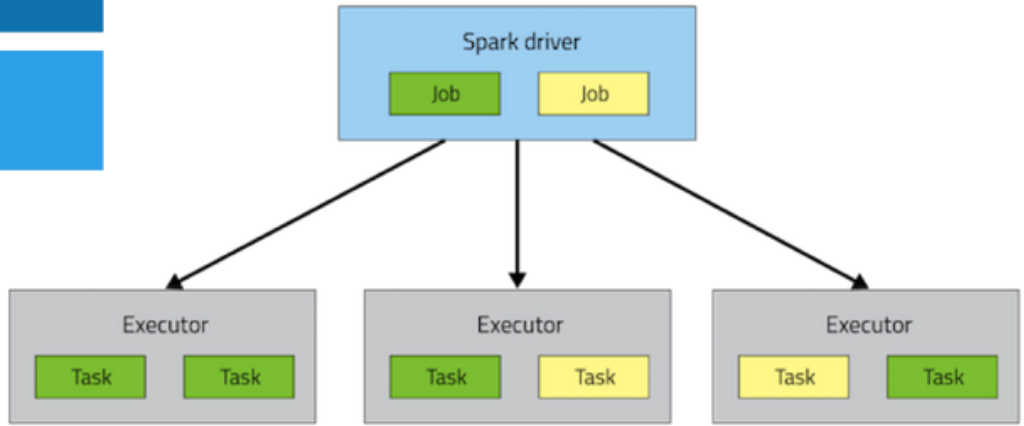
Apache Spark

Spark SQL	Spark Streaming	MLlib (machine learning)	GraphX (graph)

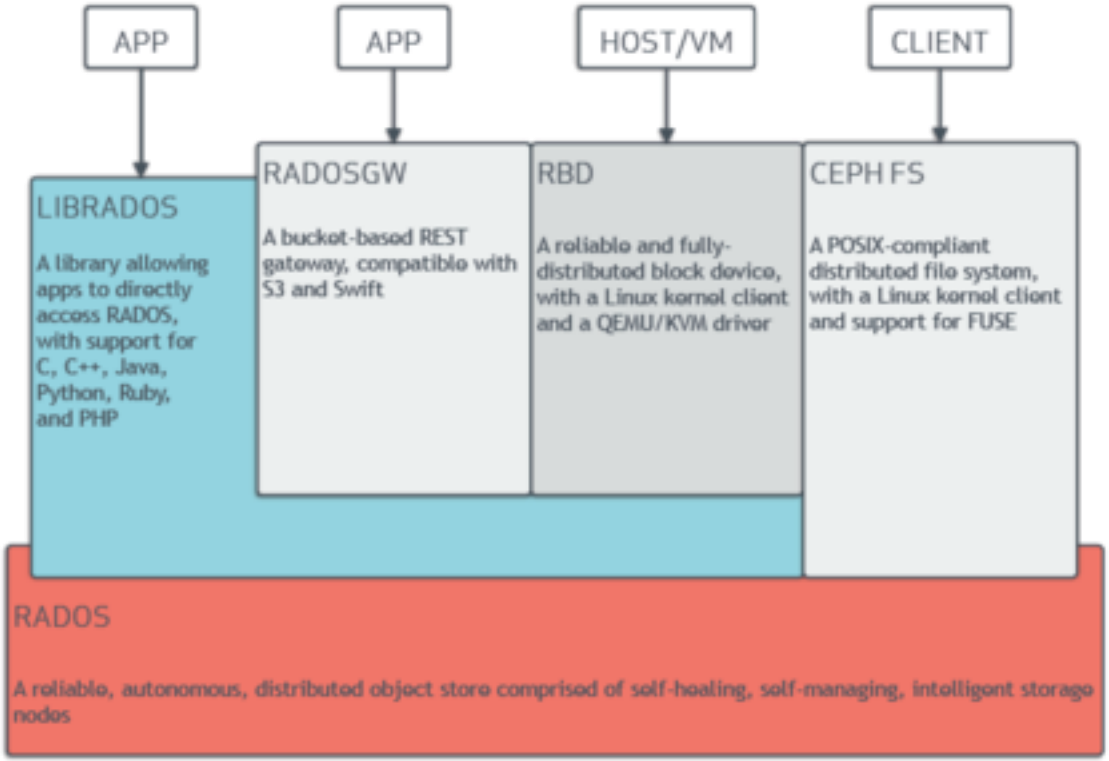
Apache Spark







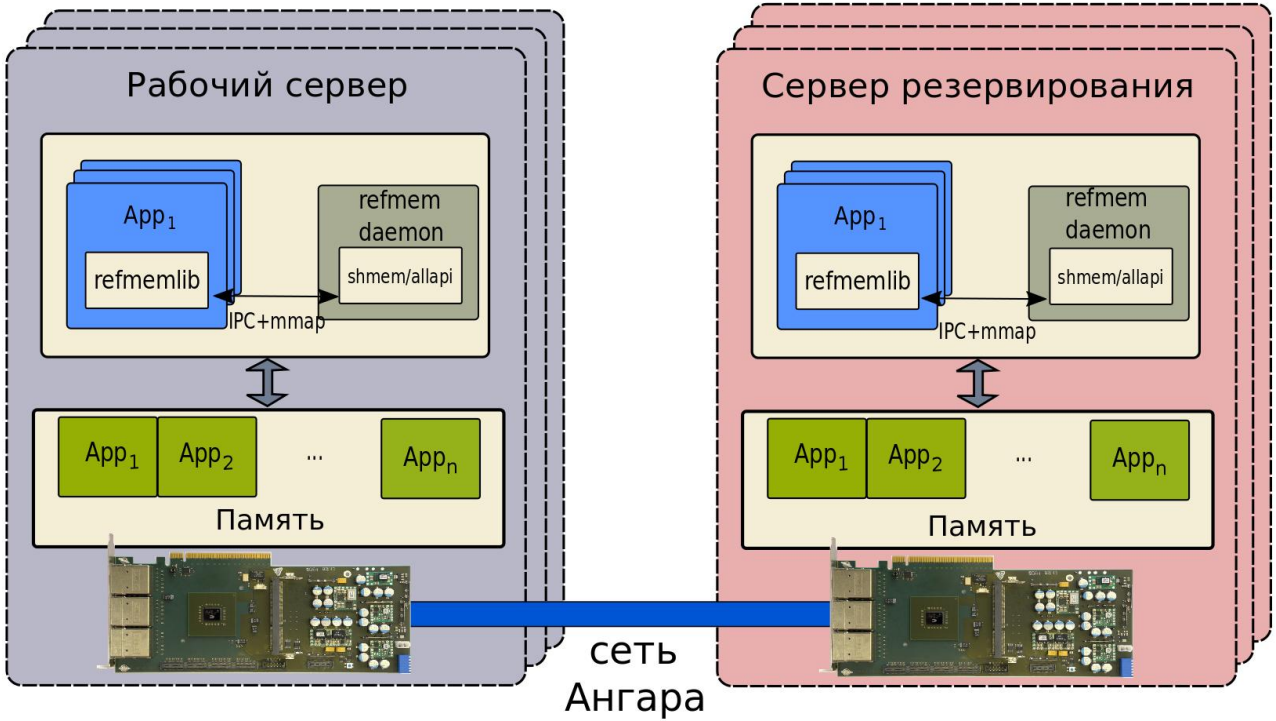
Программная система хранения данных Ceph



ceph

```
$# rados bench -p scbench 10 rand
  sec Cur ops   started  finished  avg MB/s  cur MB/s last lat(s)  avg lat(s)
    0       0         0         0         0         0           -           0
    1      16       258       242   967.695       968    0.142073   0.0610817
    2      16       487       471   941.789       916   0.0234243   0.0647762
    3      15       739       724   965.153      1012    0.145909   0.0643161
    4      15      1049      1034   1033.83      1240   0.0233676   0.0603486
    5      16      1361      1345   1075.84      1244   0.0055336   0.0579456
    6      16      1714      1698   1131.84      1412   0.0299221   0.0556169
    7      16      2065      2049    1170.7      1404    0.012719   0.0536391
    8      16      2419      2403   1201.34      1416   0.0165833   0.0523875
    9      16      2754      2738   1216.73      1340   0.0138274   0.0517339
   10      15      3103      3088   1235.04      1400   0.0764744   0.0510114
Total time run:       10.090779
Total reads made:     3104
Read size:            4194304
Object size:          4194304
Bandwidth (MB/sec):   1230.43
Average IOPS:         307
Stddev IOPS:          49
Max IOPS:             354
Min IOPS:             229
Average Latency(s):   0.0512704
Max latency(s):       0.22856
Min latency(s):       0.00462222
```

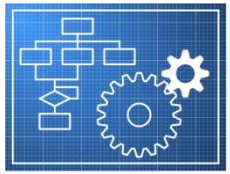
Рефлективная память



- Область памяти, которая автоматически реплицируются на другие узлы, на которых работает задача
- Возможен автоматический и программный режимы управления

Сеть Ангара: поддержка и сервисы

- Настройка программного обеспечения на вычислительных системах, в том числе MPI
- Оперативная поддержка пользователей
  angara.nicevt.ru
  support@angara.nicevt.ru



- Профилирование и адаптация прикладного ПО

Контакты:

117587, Москва, Варшавское ш, 125

angara@nicevt.ru