РФЯЦ-ВНИИТФ
РОСАТОМ

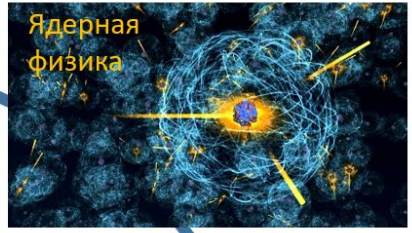
Высокопроизводительные вычисления для решения современных задач математической физики

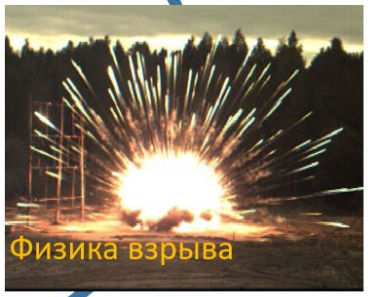
Карпеев Артём Владимирович

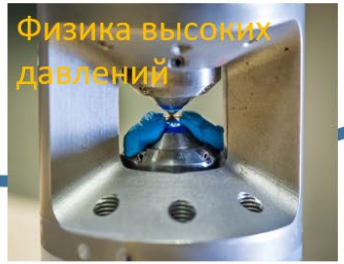
Зимняя школа по физике высоких плотностей энергии
РФЯЦ-ВНИИТФ, 28.02.2024

Фундаментальные и прикладные исследования РФЯЦ-ВНИИТФ

РФЯЦ-ВНИИТФ
РОСАТОМ

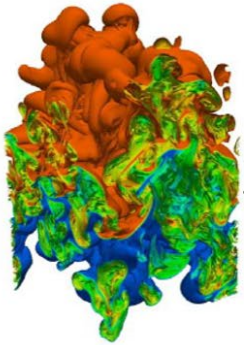






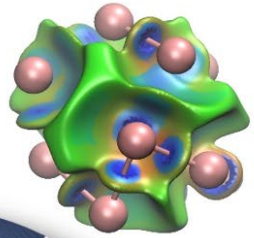
Долгосрочные комплексные задачи НТО

РФЯЦ-ВНИИТФ
РОСАТОМ

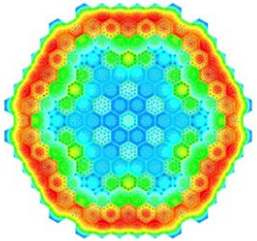


Моделирование турбулентности

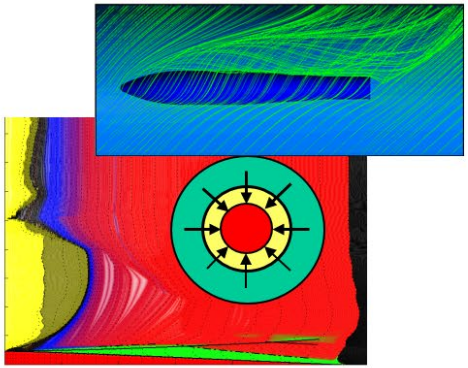
Квантовомеханические расчеты



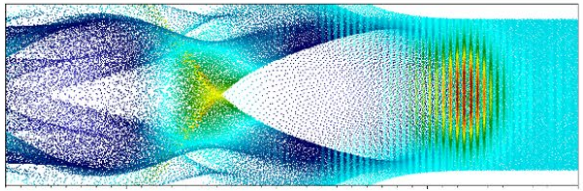
Моделирование ядерных реакций, переноса нейтронов и заряженных частиц



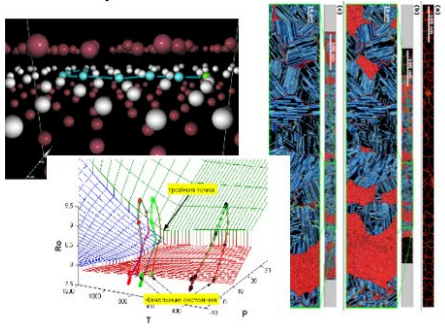
Холодная и радиационная газовая гидродинамика



Моделирование взаимодействия излучения с веществом



Компьютерные модели материалов



Что такое суперкомпьютер?

РФЯЦ-ВНИИТФ
РОСАТОМ

Суперкомпьютер – это вычислительная система, производительность которой во много раз превосходит производительность современных ей компьютеров массового выпуска

Суперкомпьютер — изделие мелкосерийное или даже штучное. При его производстве применяются технологии и подходы высокой стоимости.

компьютер массового выпуска

суперкомпьютер

Измерение производительности

РФЯЦ-ВНИИТФ
РОСАТОМ

FLOPS – FIoating point **Op**erations per **S**econd – количество операций с плавающей запятой за секунду

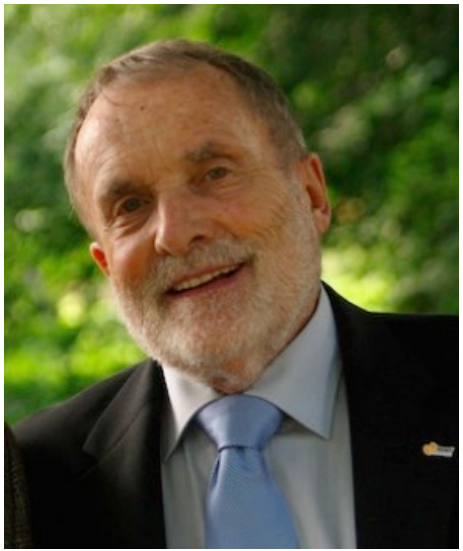
UNIT = 10**6 TIME/(1/3 100**3 + 100**2)

Facility	TIME N=100 secs.	UNIT micro-secs.	Computer	Type	Compiler
NCAR	.049	0.14	CRAY-1	S	CFT, Assembly BLAS
LASL	.148	0.43	CDC 7600	S	FTN, Assembly BLAS
NCAR	.192	0.56	CRAY-1	S	CFT
LASL	.210	0.61	CDC 7600	S	FTN
Argonne	.297	0.86	IBM 370/195	D	H
NCAR	.359	1.05	CDC 7600	S	Local
Argonne	.388	1.33	IBM 3033	D	H
NASA Langley	.489	1.42	CDC Cyber 175	S	FTN
U. Ill. Urbana	.506	1.47	CDC Cyber 175	S	Ext. 4.6
LLL	.554	1.61	CDC 7600	S	CHAT, No optimize
SLAC	.579	1.69	IBM 370/168	D	H Ext., Fast mult.
Michigan	.631	1.84	Amdahl 470/V6	D	H
Toronto	.890	2.59	IBM 370/165	D	H Ext., Fast mult.
Northwestern	1.44	4.20	CDC 6600	S	FTN
Texas	1.93*	5.63	CDC 6600	S	RUN
China Lake	1.95*	5.69	Univac 1110	S	V
Yale	2.59	7.53	DEC KL-20	S	F20
Bell Labs	3.46	10.1	Honeywell 6080	S	Y
Wisconsin	3.49	10.1	Univac 1110	S	V
Iowa State	3.54	10.2	Itel AS/5 mod3	D	H
U. Ill. Chicago	4.10	11.9	IBM 370/158	D	G1
Purdue	5.69	16.6	CDC 6500	S	FUN
U. C. San Diego	13.1	38.2	Burroughs 6700	S	H
Yale	17.1*	49.9	DEC KA-10	S	F40

* TIME(100) = (100/75)**3 SGEFA(75) + (100/75)**2 SGESL(75)

LINPACK [1977] → … → HPL [**H**igh **P**erformance **L**inpack, 2024]

Top 500 – рейтинг лучших



Hans W. Meuer

немецкий учёный в области вычислительной техники

Ганс Майер в 1983 начал вести рейтинг суперкомпьютеров по всему миру, ранжируя их по пиковой производительности

В 1993 Майер и Донгарра объединили усилия и создали рейтинг

TOP 500 The List.

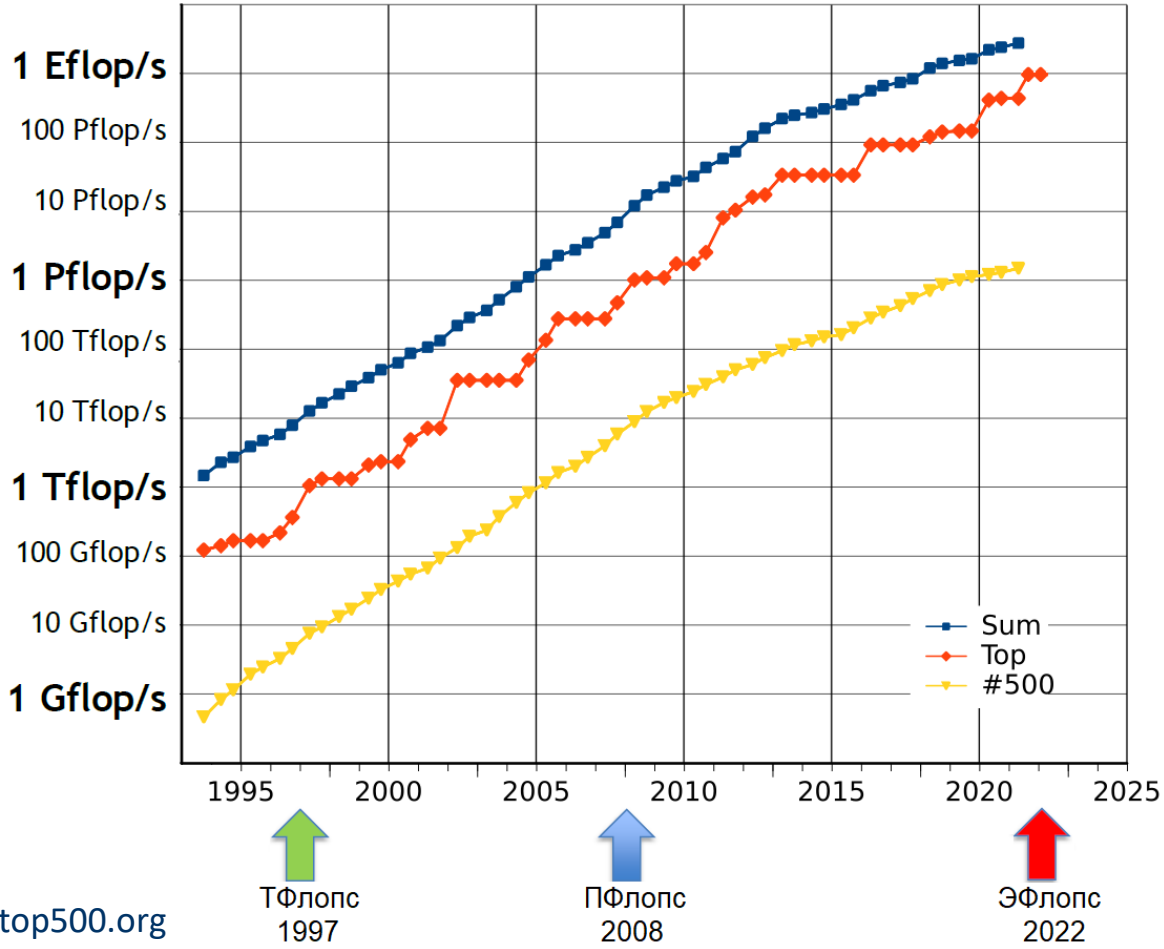
Jack Dongarra

американский учёный в области вычислительной техники

https://top500.org

Эволюция в суперкомпьютерах

РФЯЦ-ВНИИТФ
РОСАТОМ



RoadRunner [LANL]
• 1ПФ, Top-2008 • 6k CPU+12k GPU • S = 1100 м^2 • E = 4 МВт • 130 M$ • Списан в 2013

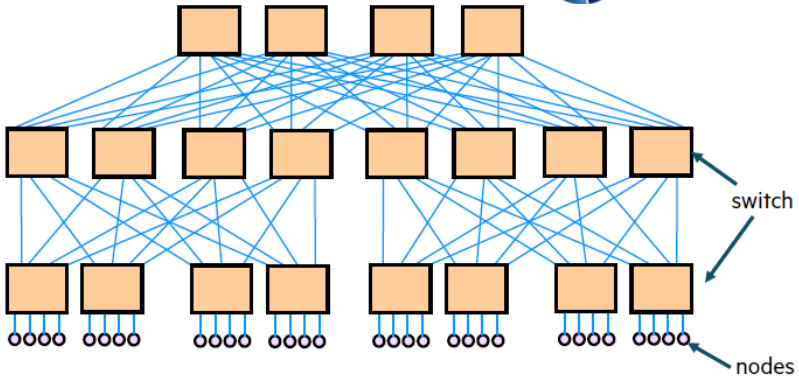
https://top500.org

Топологии объединения узлов

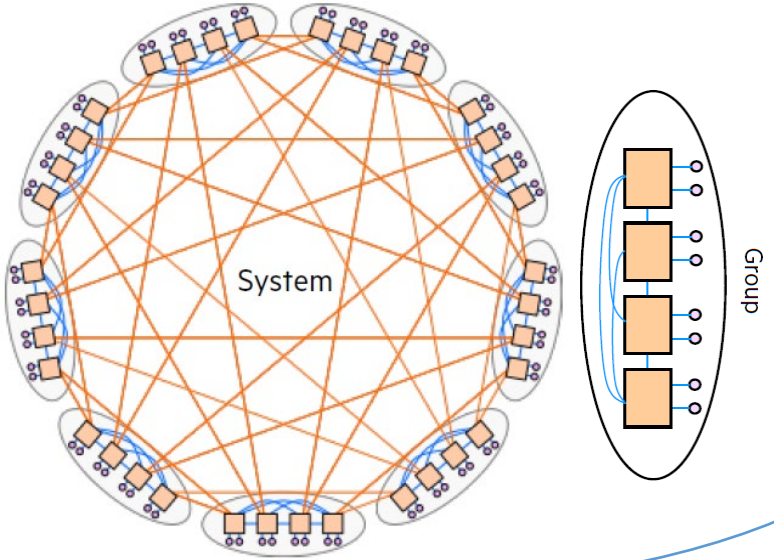
РФЯЦ-ВНИИТФ
РОСАТОМ

- Квадрат - коммутатор
- Кружок - узел

- Полносвязанное FatTree:
  - одинаковая Bw на всех уровнях
  - неблокирующие парные обмены для всех узлов
  - высокая стоимость масштабирования



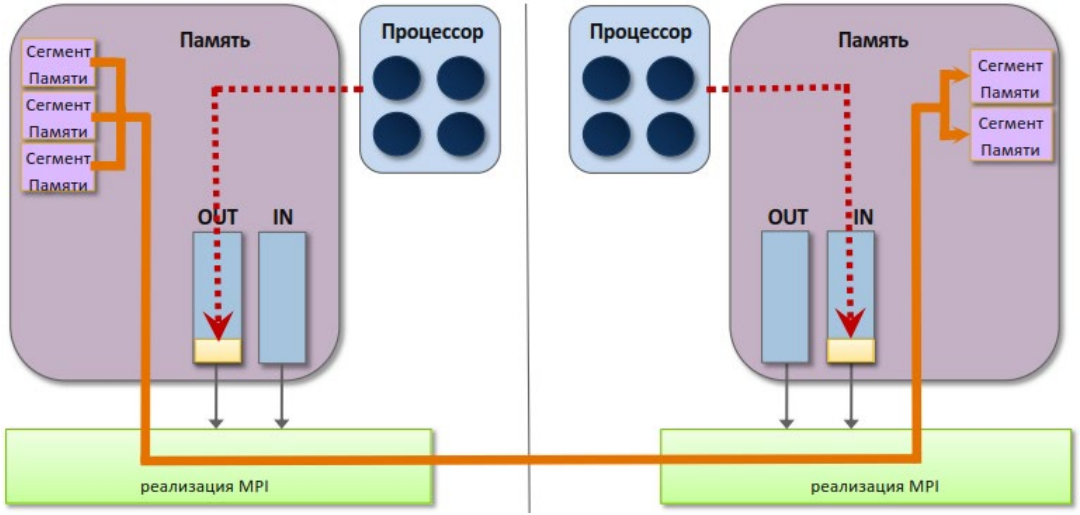
- Dragonfly [Frontier, Aurora, ElCapitan]
  - попарная связь всех коммутаторов группы
  - все группы имеют попарную связь
  - различие Bw по уровням
  - экономия на коммутаторах и кабелях



Распределённая память – MPI

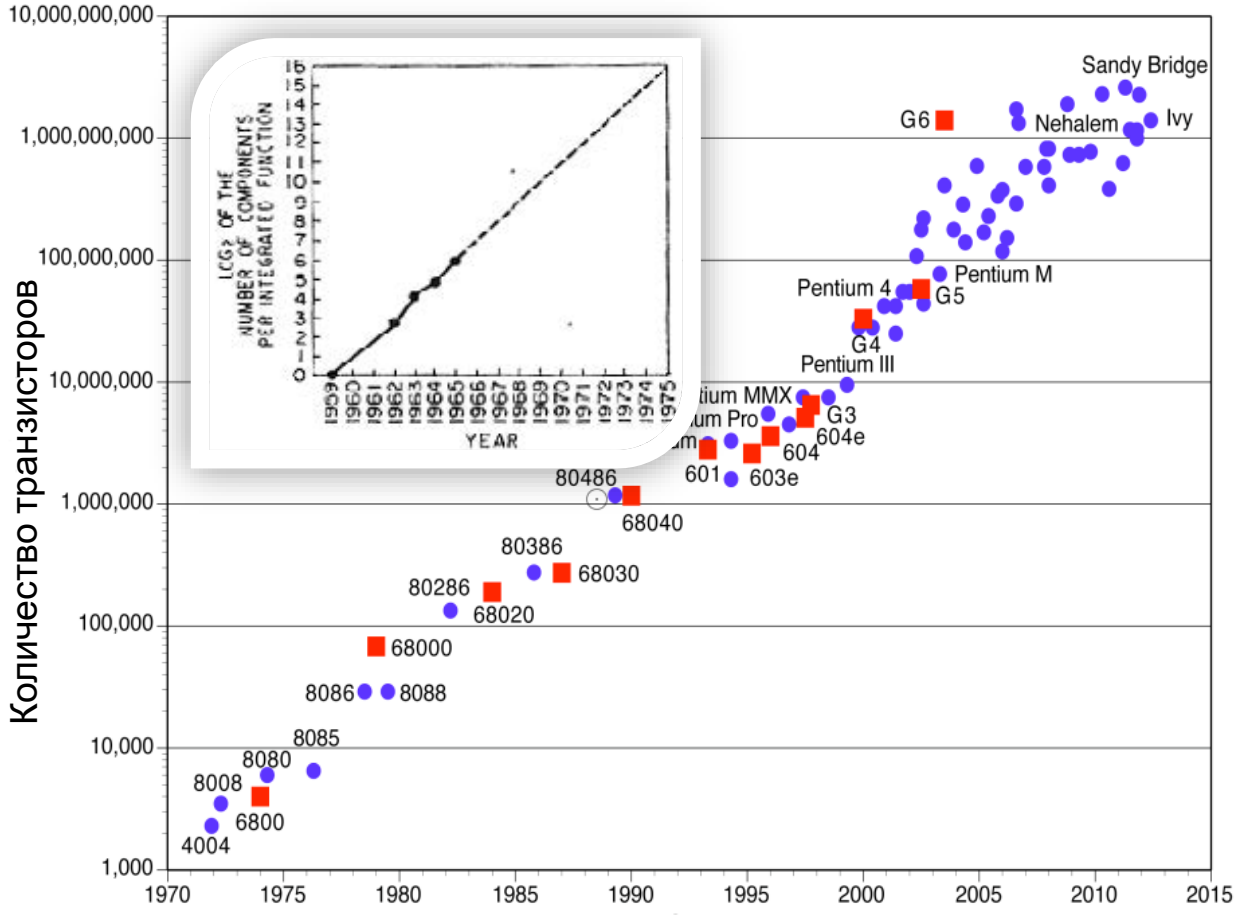
MPI [**M**essage **P**assing **I**nterface] – «стандарт» организации передачи информации между процессами, выполняющими общую задачу

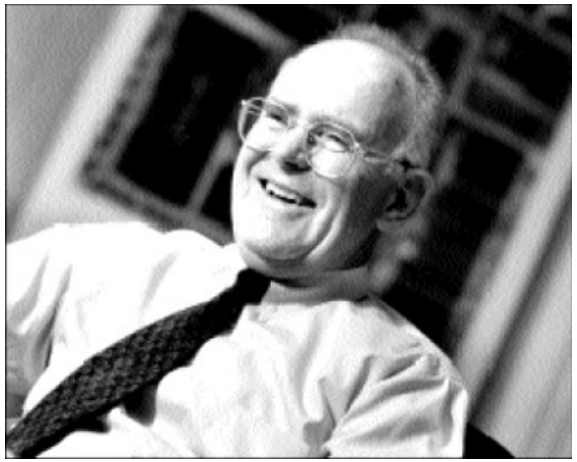
MPI-1 [1994], MPI-2 [1997], MPI-3 [2012]



http://www.mpi-forum.org

Закон Мура





Gordon E. Moore в 1965 году:

«Количество транзисторов, размещаемых на кристалле интегральной схемы, удваивается каждые 18 месяцев»

G. Moore, Electronic Magazine, 1965

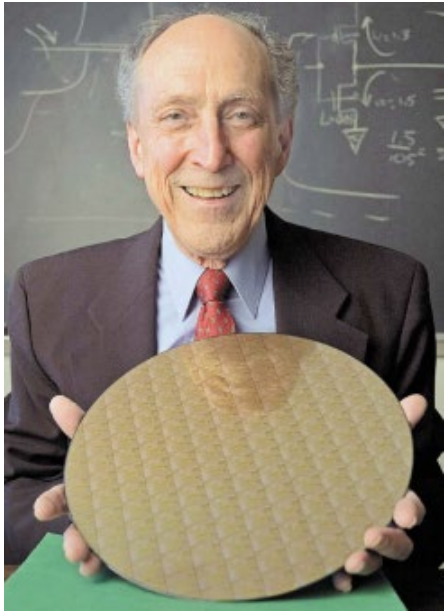
Теория масштабирования Деннарда

РФЯЦ-ВНИИТФ
РОСАТОМ

Роберт Деннард в 1974 году объяснил закон Мура:

«*Уменьшая размеры транзистора и повышая тактовую частоту процессора, мы можем легко повышать его производительность*»

Эмпирический закон соотнёс масштабирование с производительностью и объяснил, каким именно способом следует двигаться в направлении ее повышения.



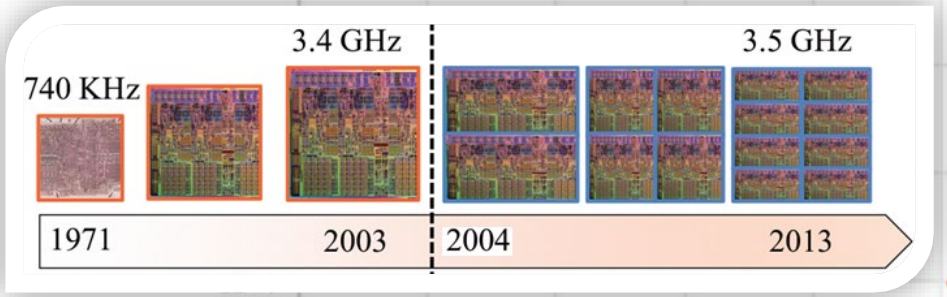
Robert H. Dennard

в 1968 году изобрёл динамическую память с произвольным доступом (DRAM)

Dennard et al., IEEE JSSC, 1974

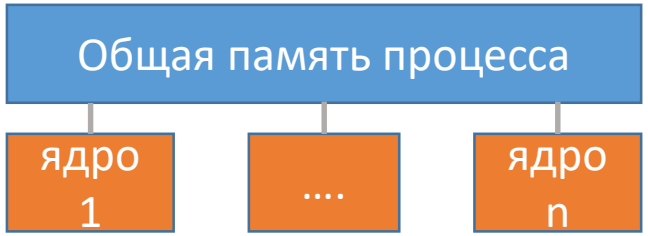
Многоядерность

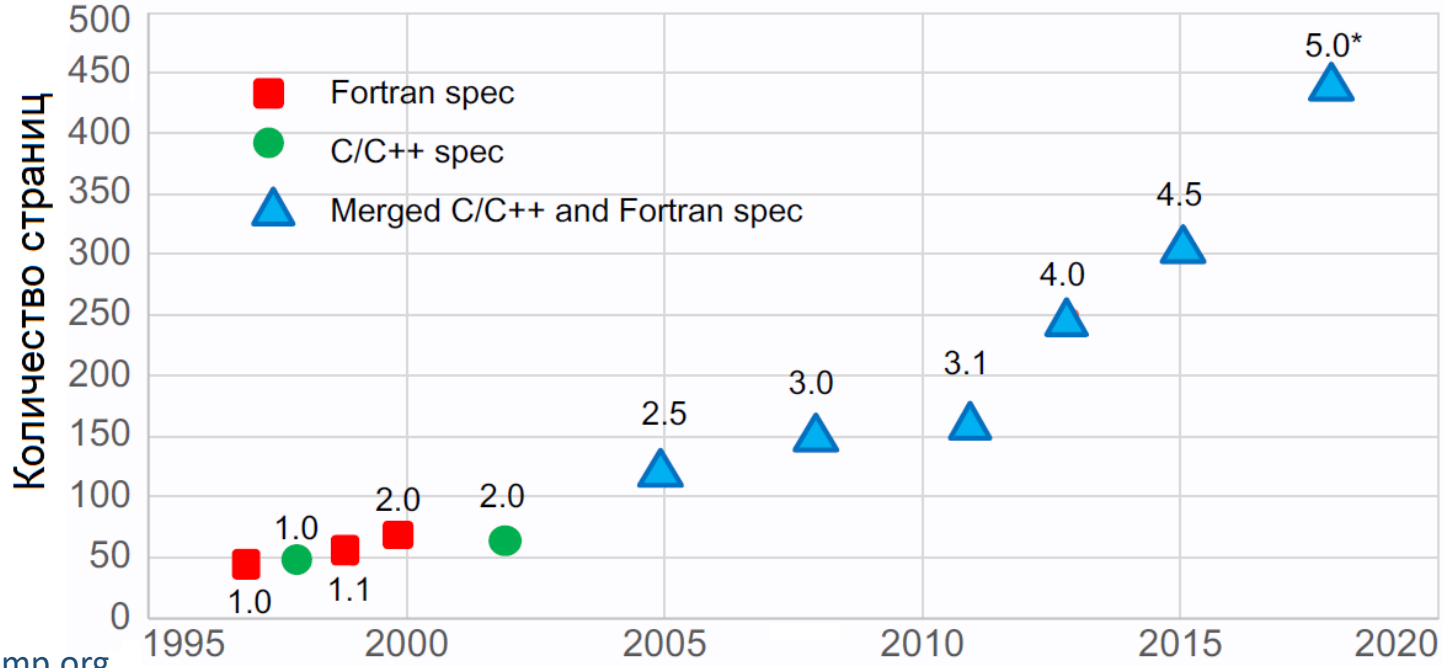
РФЯЦ-ВНИИТФ
РОСАТОМ



Общая память – OpenMP

OpenMP [**Open M**ulti-**P**rocessing] – открытый стандарт для распараллеливания программ на общей памяти





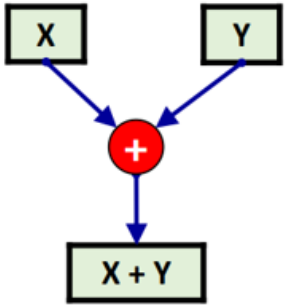
http://www.openmp.org

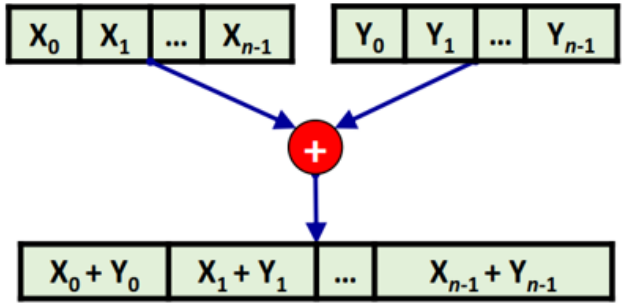
Процессоры с векторным расширением

РФЯЦ-ВНИИТФ
РОСАТОМ

$$\text{FLOPS} = Q_{ядер} \cdot F \cdot \frac{FLOP}{такт},$$

F – тактовая частота





SIMD [**S**ingle **I**nstruction **M**ultiple **D**ata]
параллелизм на уровне данных

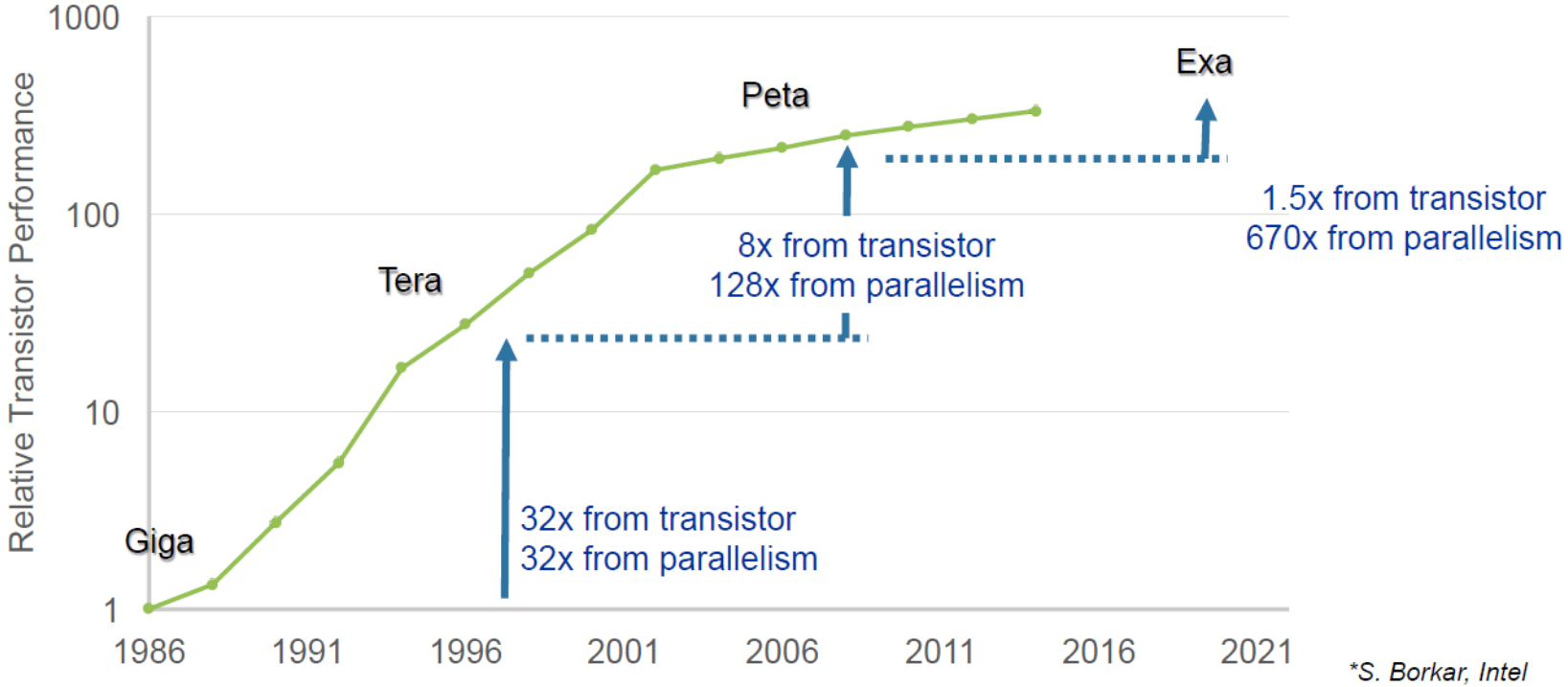
```
add Z, X, Y
```

```
add.v Z[0:n-1], X[0:n-1], Y[0:n-1]
```

Процессоры Intel	Год	Набор SIMD	FLOP/такт
Core2	2007	SSE2	2
Nehalem	2009	SSE4	4
SandyBridge	2011	AVX	8
Haswell	2013	AVX 2	16
Skylake	2015	AVX 512	32

В чем основной вызов для софта?

РФЯЦ-ВНИИТФ
РОСАТОМ

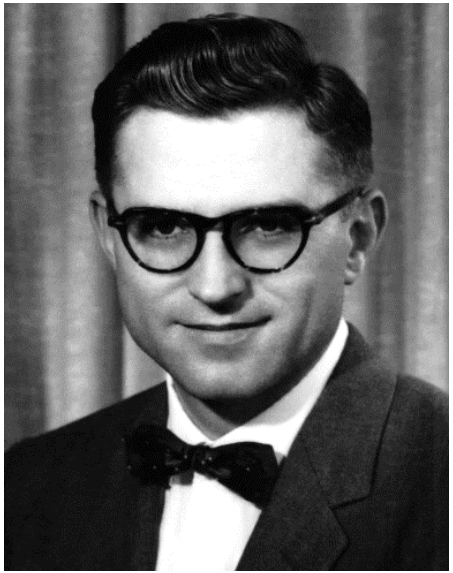


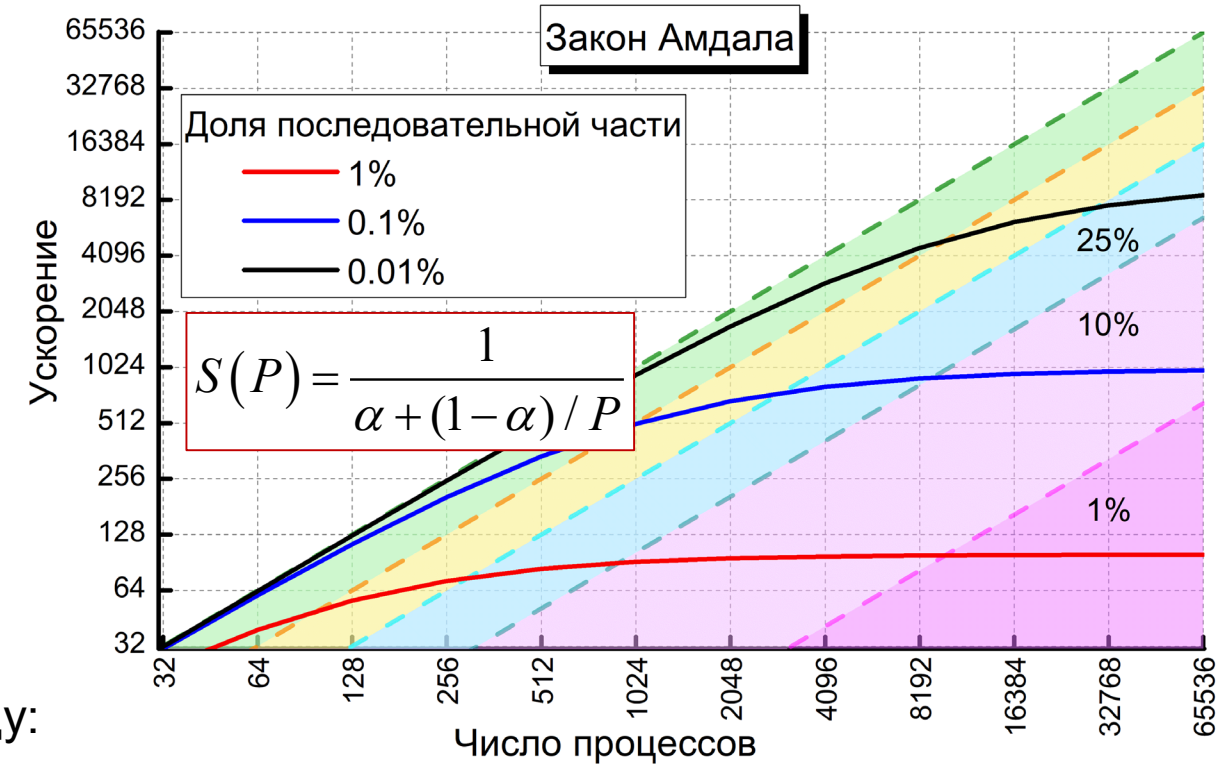
*S. Borkar, Intel

Весь прирост производительности экза-ВВС достигается за счёт **параллелизма**, причем большая его часть будет внутри **узла**

Закон Амдала – предел ускорения

РФЯЦ-ВНИИТФ
РОСАТОМ





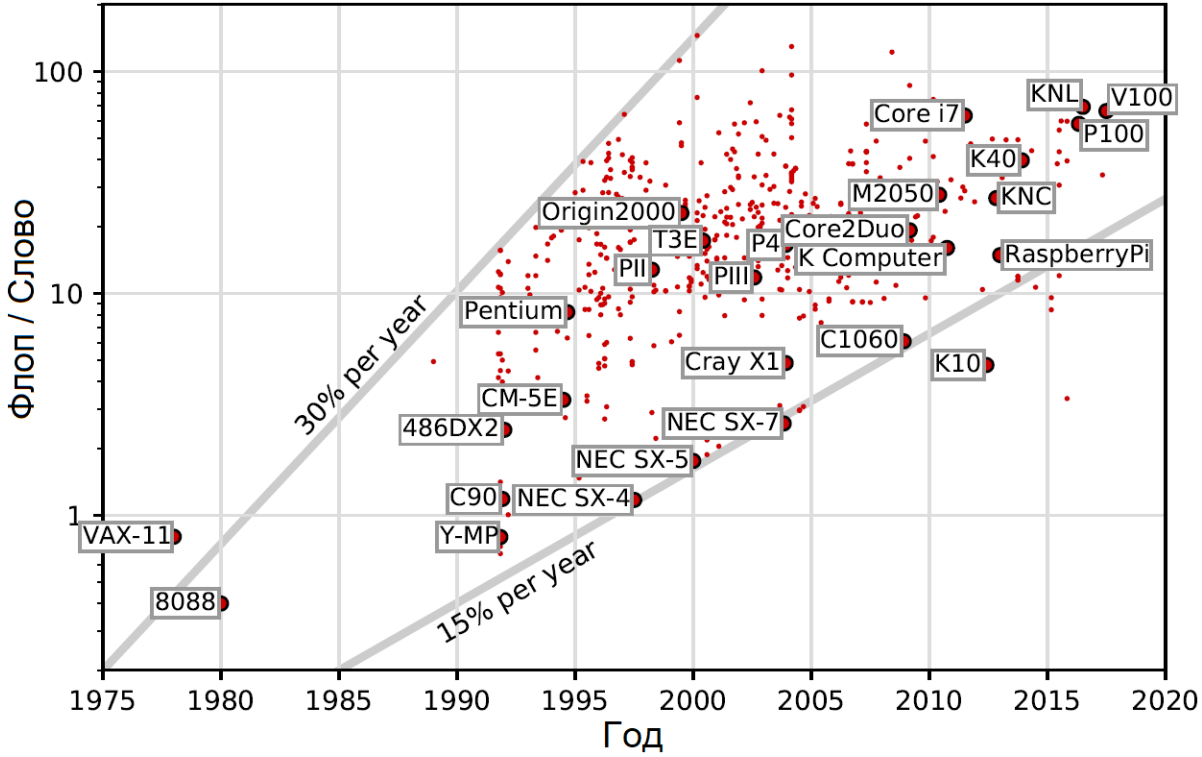
Джин Амдал в 1967 году:

«В случае, когда задача разделяется на несколько частей, суммарное время её выполнения на параллельной системе не может быть меньше времени выполнения самого медленного фрагмента»

«Стена памяти»

РФЯЦ-ВНИИТФ
РОСАТОМ

Производительность вычислений / производительность памяти



Год	Вычислитель	Флоп/Слово
1980	8088	**0.2**
2020	V100	**70**

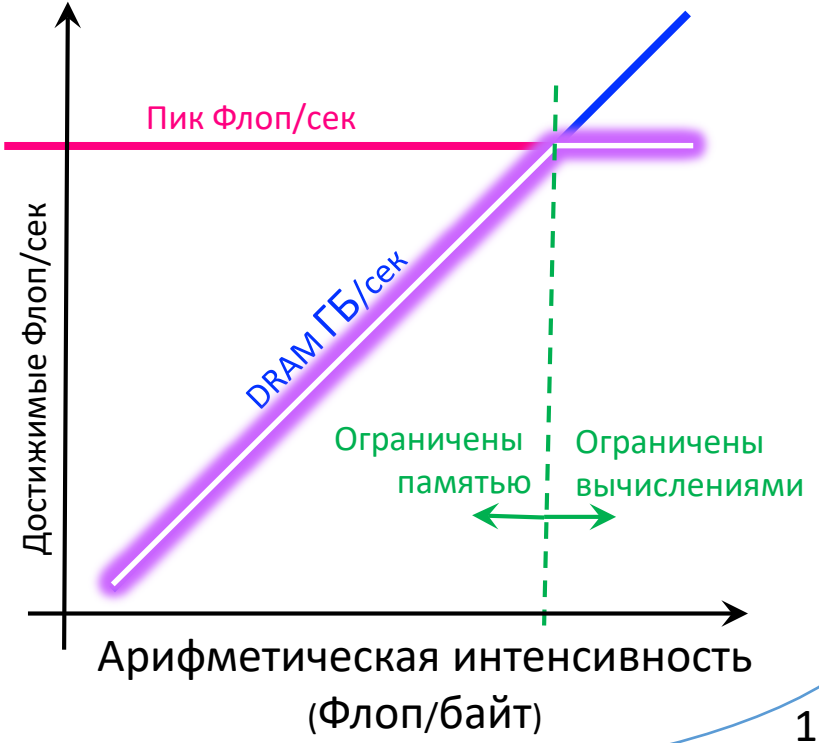
Относительный рост производительности за 35 лет **>100 раз!!!**

Roofline-модель

Мера производительности (FLOP/s) как функция арифметической интенсивности (т.е. количество FLOPs на байт перемещенных из DRAM/cache данных)

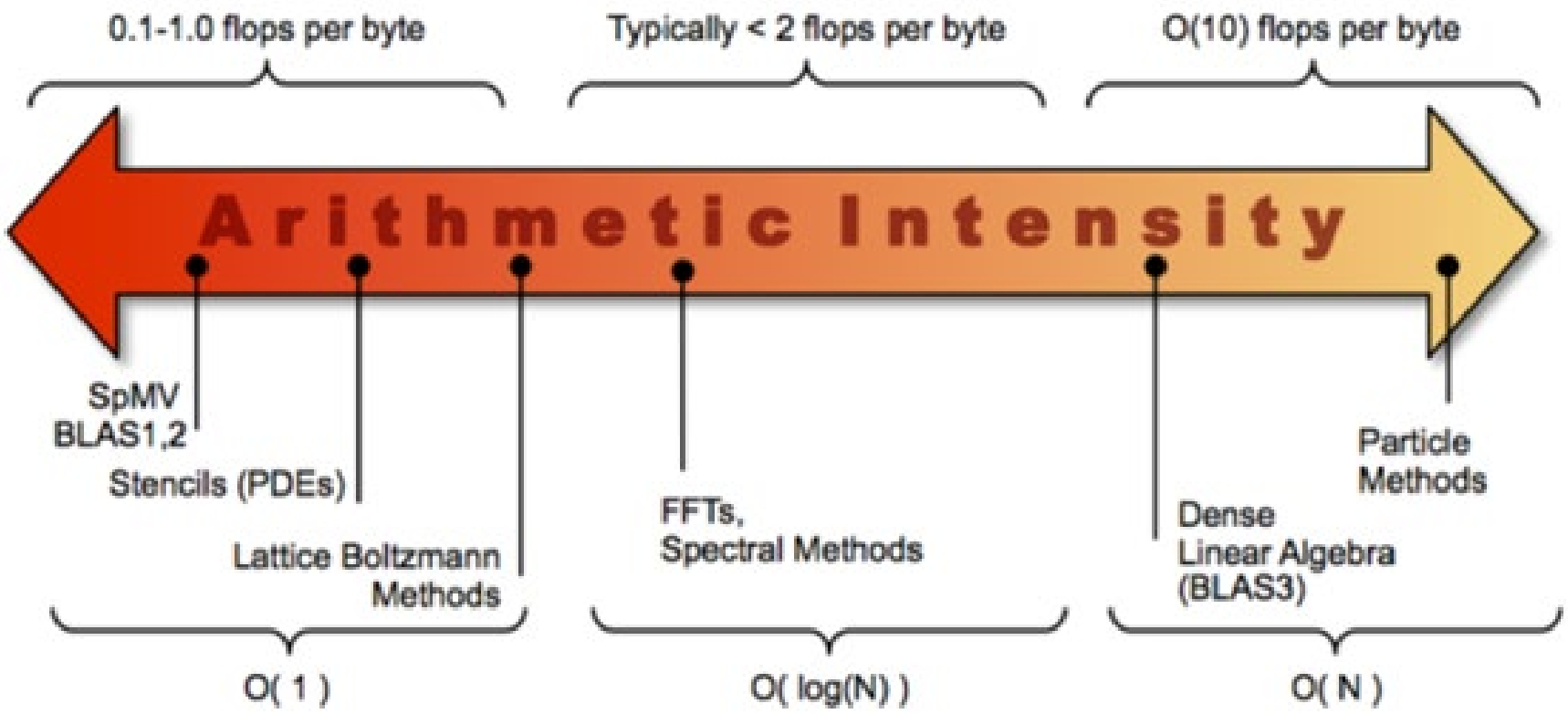
Производительность ограничена:

- производительностью вычислителя
- пропускной способностью памяти



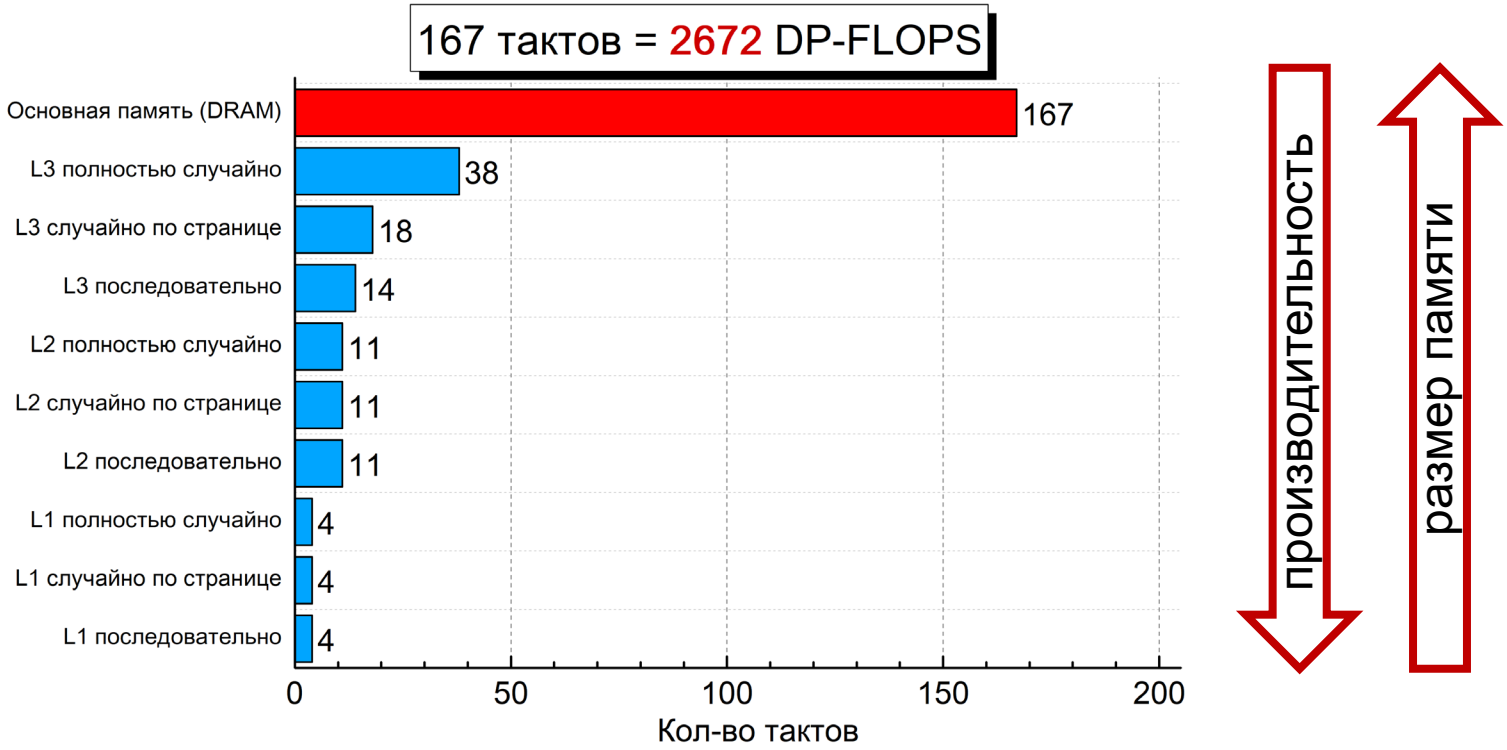
S. Williams et al., Comm. ACM 52 (2009)

Классификация численных методов



https://crd.lbl.gov/departments/computer-science/PAR/research/roofline

Иерархическая память

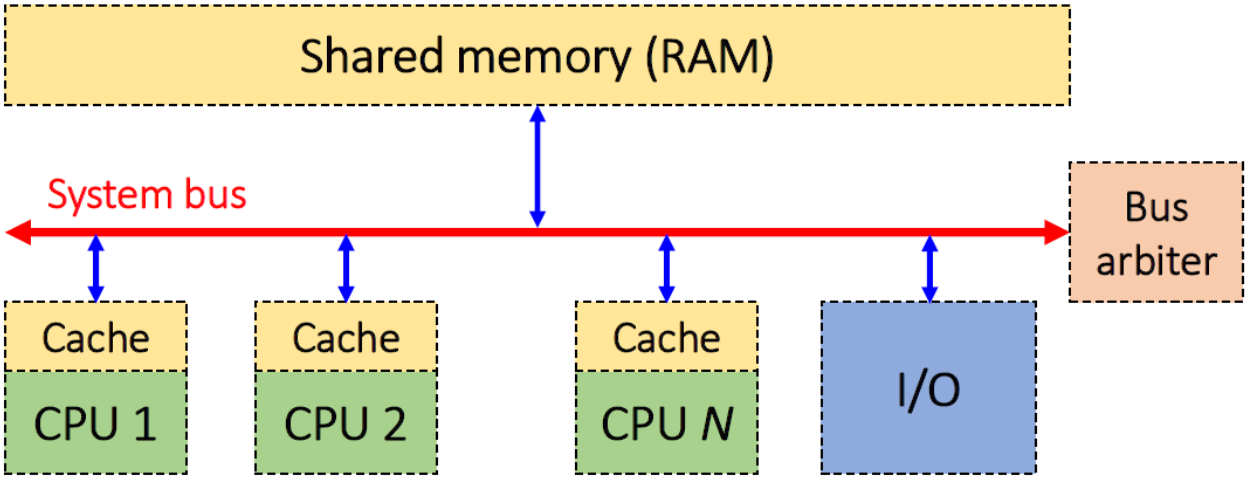


Доступ к данным в L3 в **10 раз** быстрее чем в ОП

J. Dongarra, A Look at HPC, RSC days 2015

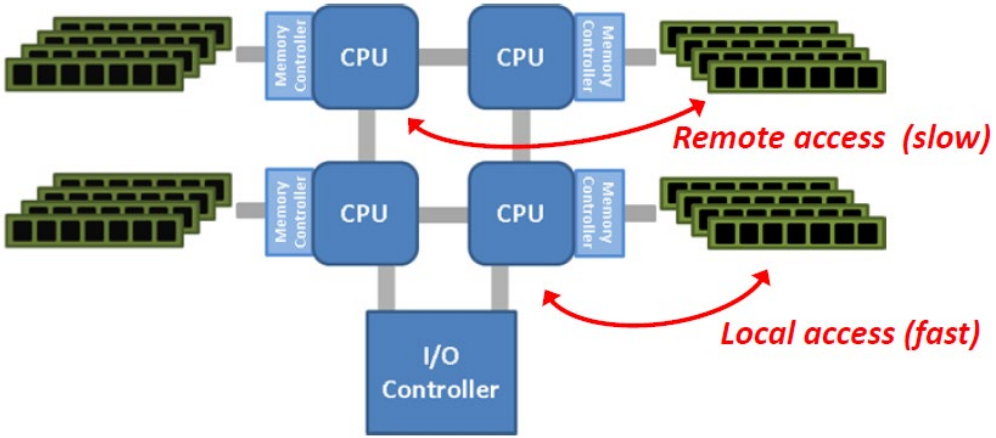
SMP-система

- **SMP** (**S**ymmetric **M**ulti**P**rocessing) – архитектура вычислительной системы с равномерным (симметричным) доступом к разделяемой общей памяти
- Системная шина (System bus) – узкое место, ограничивающее масштабируемость узла



NUMA-система

- **NUMA** (**N**on-**U**niform **M**emory **A**rchitecture) – архитектура вычислительной системы с неравномерным доступом к разделяемой общей памяти
- Ядра сгруппированы в NUMA-острова со своей локальной памятью
- Доступ к локальной памяти NUMA-острова занимает существенно меньше времени (**2-3 раза**), чем к памяти соседнего NUMA-острова

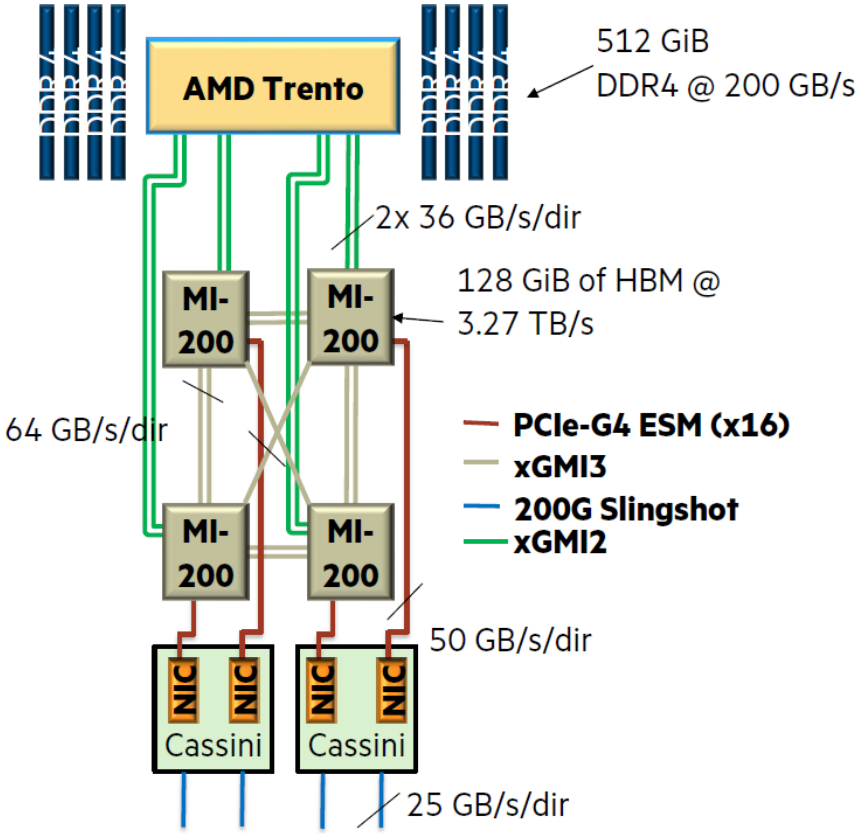


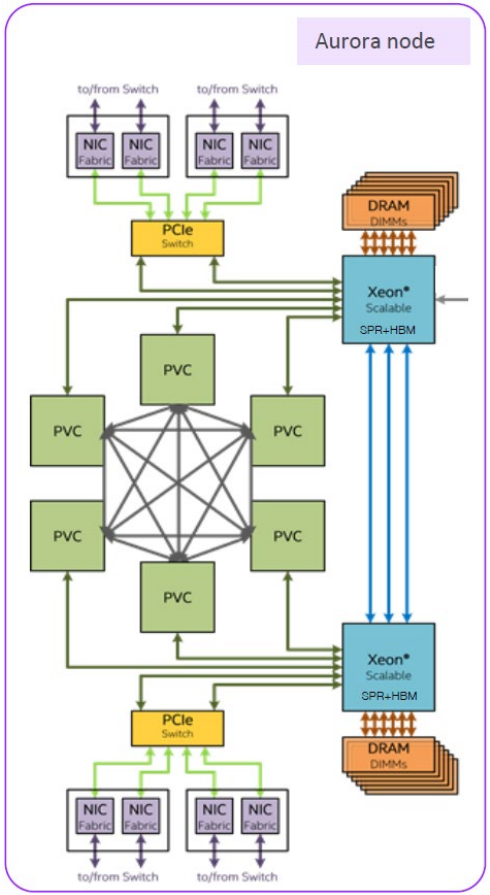
OPT 2008	1.6
HSW 2012	2.0
EPYC 2020	3.0

Откуда в супервычислениях GPU?

- **GPU** (**G**raphics **P**rocessing **U**nit) – адаптированный под работу с графикой процессор
- Б**о**льшая часть площади GPU занята элементарными модулями (ALU, FPU, Load, Store)
- Устройство управления (Control) – существенно проще чем у CPU

Архитектура гибридного узла





Frontier, Underwood (HPE), ATPESC 2022

Aurora, Muralidharan (ALCF), ATPESC 2022

Top 500 сегодня

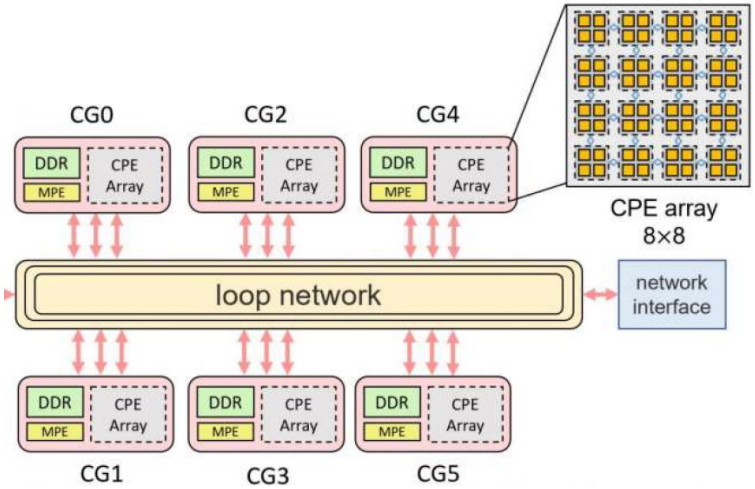
№	Имя	#узлов	#ядер	Узел	Q, ПФ	HPL, ПФ	HPCG, ПФ	P, МВт
1	Frontier	9.2k	8.7M	CPU+4GPU	1680	1194 / 71%	14.1 / **0.8%**	23
2	Aurora	5.5k	4.7M	2CPU+6GPU	1059	585 / 55%		25
3	Eagle	3.6k	7.6M	CPU+4GPU	847	561 / 66%		9
4	Fugaku	**159k**	7.6M	CPU	537	442 / 82%	**16.0** / **3%**	30
5	LUMI	2.9k	2.2M	CPU+4GPU	429	309 / 72%	3.4 / **0.8%**	7

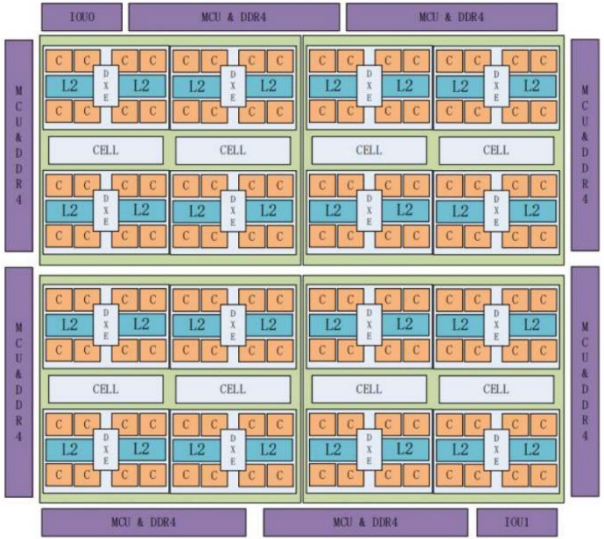
- 10 ВВС > 100ПФ, 77 ВВС > 10ПФ, 500-й > 2ПФ
- Использовано 3 типа узлов: CPU; **CPU+GPU**; на «лёгких ядрах» (ShenWei, ARM, KNL)
- ВК с гибридными узлами – основной тренд развития ВУ
- Интерес к супервычислениям растёт по всему миру во всех областях (>30% из TOP500 – промышленность)
- Exascale проекты: США, Китай, Европа

Возможно экза-эра наступила в 2021…

РФЯЦ-ВНИИТФ
РОСАТОМ

OceanLight 1.3EF, 36МВт, ~100k-узлов:
RISC-ShenWei 6*(64+1)ядер, 2.25ГГц,
6*16=96ГБ(DDR4) 307ГБ/с, сетевой адаптер
(NIC) 2*PCIe4 [4*56Гб/сек] ~500Гб/сек, 14TF

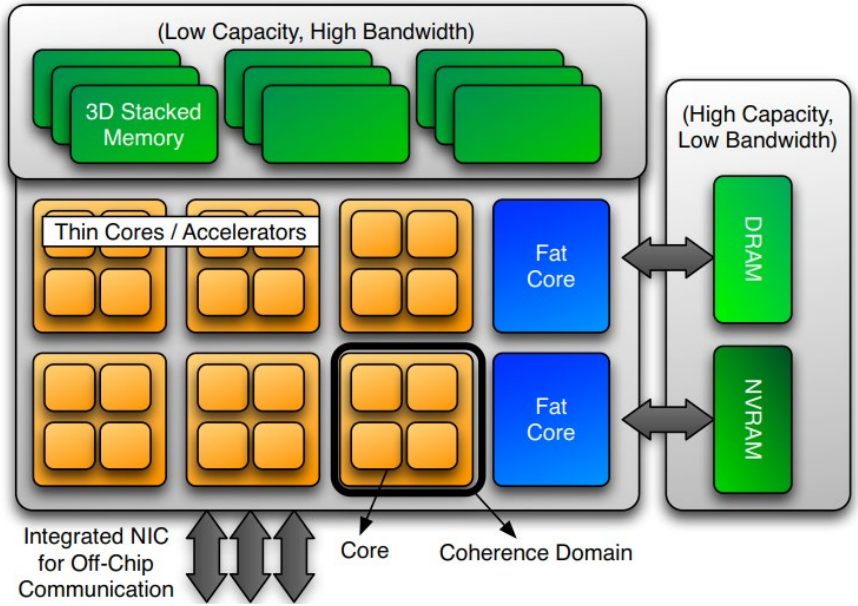




Tianhe-3 1.7EF: ARM-CPU(2+TF)+x*Matrix3000,
128+RISC-ядер 2ГГц, 10+TF, 8*DDR4/CPU =
204.8ГБ/с

2025: 2 ВВС **10EF!**

«Очевидный прогноз»



	AMD MI300A
Fat Core	24 “Zen4” x86
Thin Core	228 GPU CU
HBM	128 ГБ
HBM Bw	5.3 ТБ/с
DRAM	> 4 ТБ
NIC	Infinity Fabric 8*128 ГБ/с
Perf FP64vec	61.3 ТФ
Max TDP	550-760 Вт

“The Scientific Grand Challenges Workshop: Architectures and Technology for Extreme Scale Computing” (2009)

Заключение

РФЯЦ-ВНИИТФ
РОСАТОМ

- Суперкомпьютерное моделирование – одно из наиболее активно развивающихся направлений современной науки.
- **Самые мощные** современные **суперкомпьютеры** имеют **сложные архитектуру** и **средства программирования:** ожидается, что это сложность будет только возрастать.
- **Научные и инженерные приложения**, разработанные для таких систем, также имеют достаточно **сложную структуру.**
- Требуется большое число специалистов высокого класса, которые будут двигать высокопроизводительные параллельные вычисления вперед.
- РФЯЦ-ВНИИТФ – одно из лучших мест в стране, где можно быть на «переднем крае» супервычислений.