

КОМПЬЮТЕРНОЕ И ТЕЛЕКОММУНИКАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДЕТЕКТОРА ДЛЯ СУПЕР С-ТАУ ФАБРИКИ

А.Зайцев

A.S.Zaytsev@inp.nsk.su



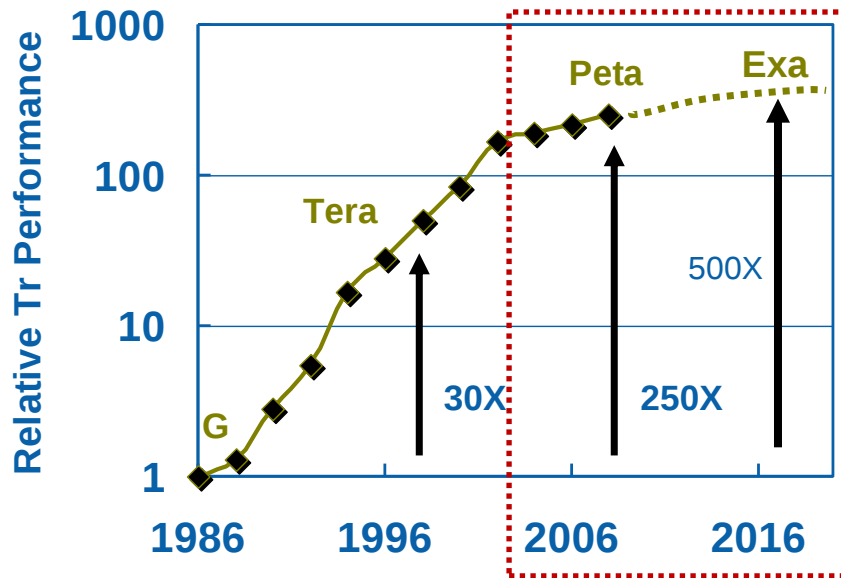
Эволюция информационных технологий в период проектирования, строительства и эксплуатации детектора

2010-2020

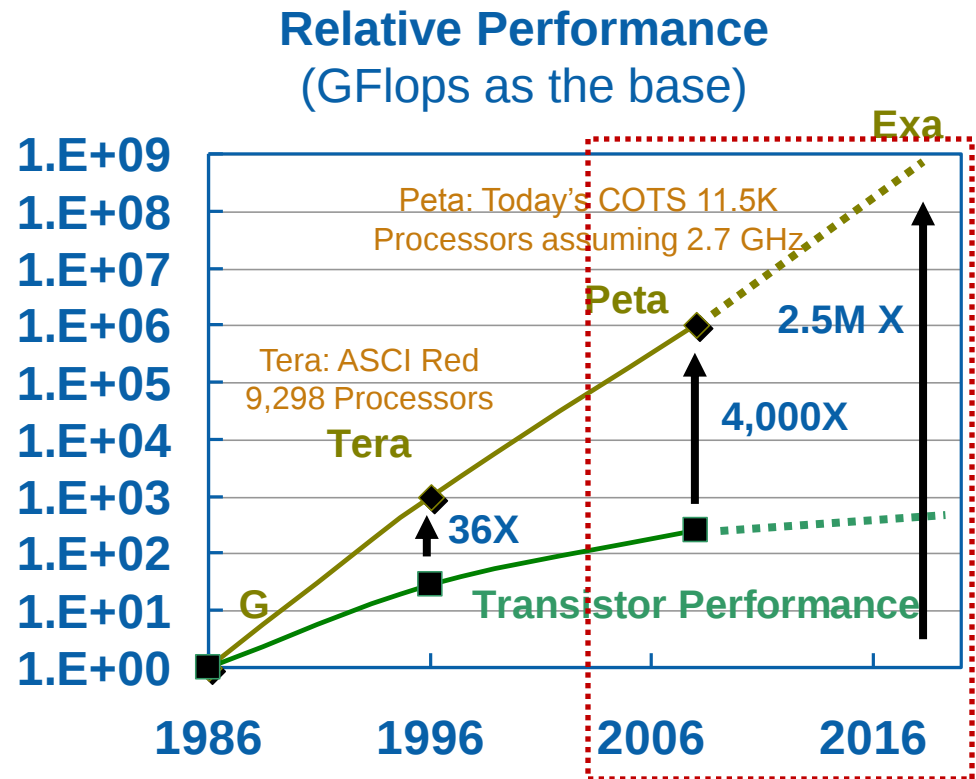
Вычислительные системы общего назначения

CPU

Moore's Law and High Performance Computing



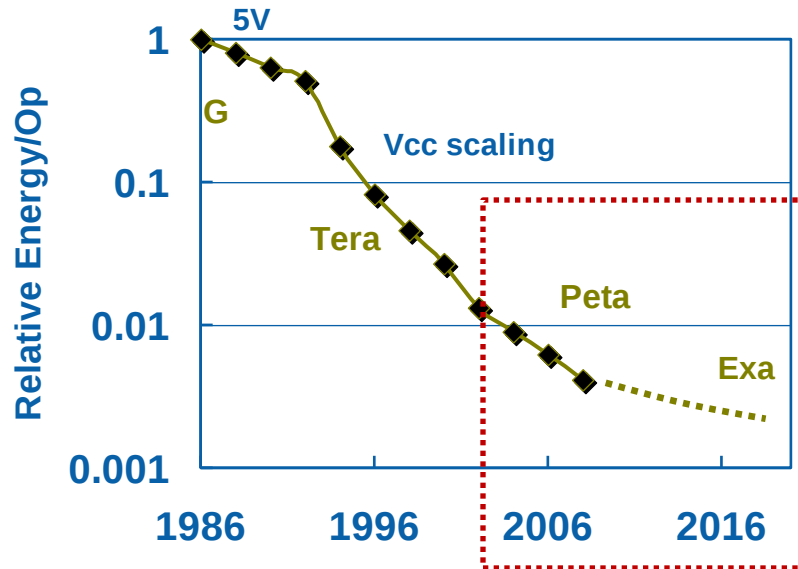
Source: Intel labs



From Peta to Exa, 2X Transistor Performance, Requiring ~30K cores @2800 SPI2K



Exponential Power and Computing Growth

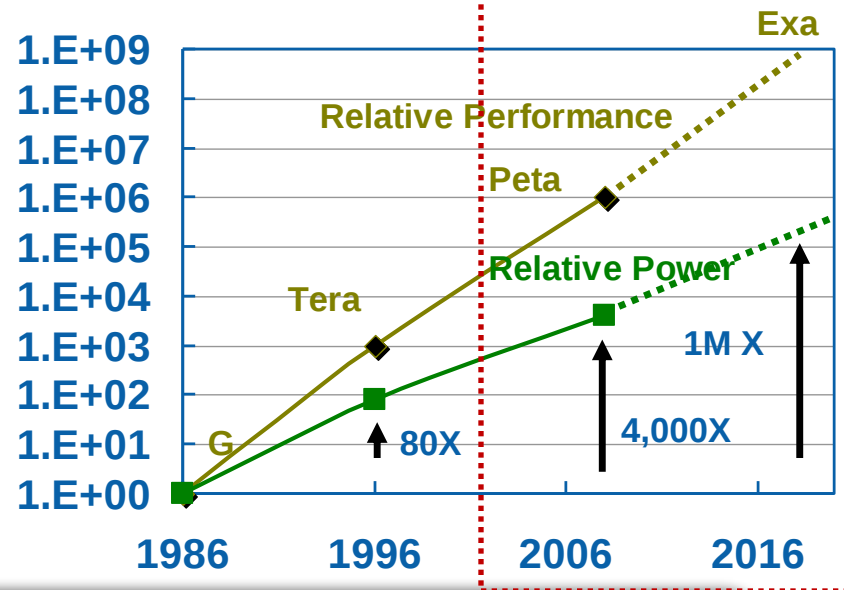


Relative Performance and Power
(GFlops as the base)

Power at a glance:

(assume 31% CPU Power in a system)

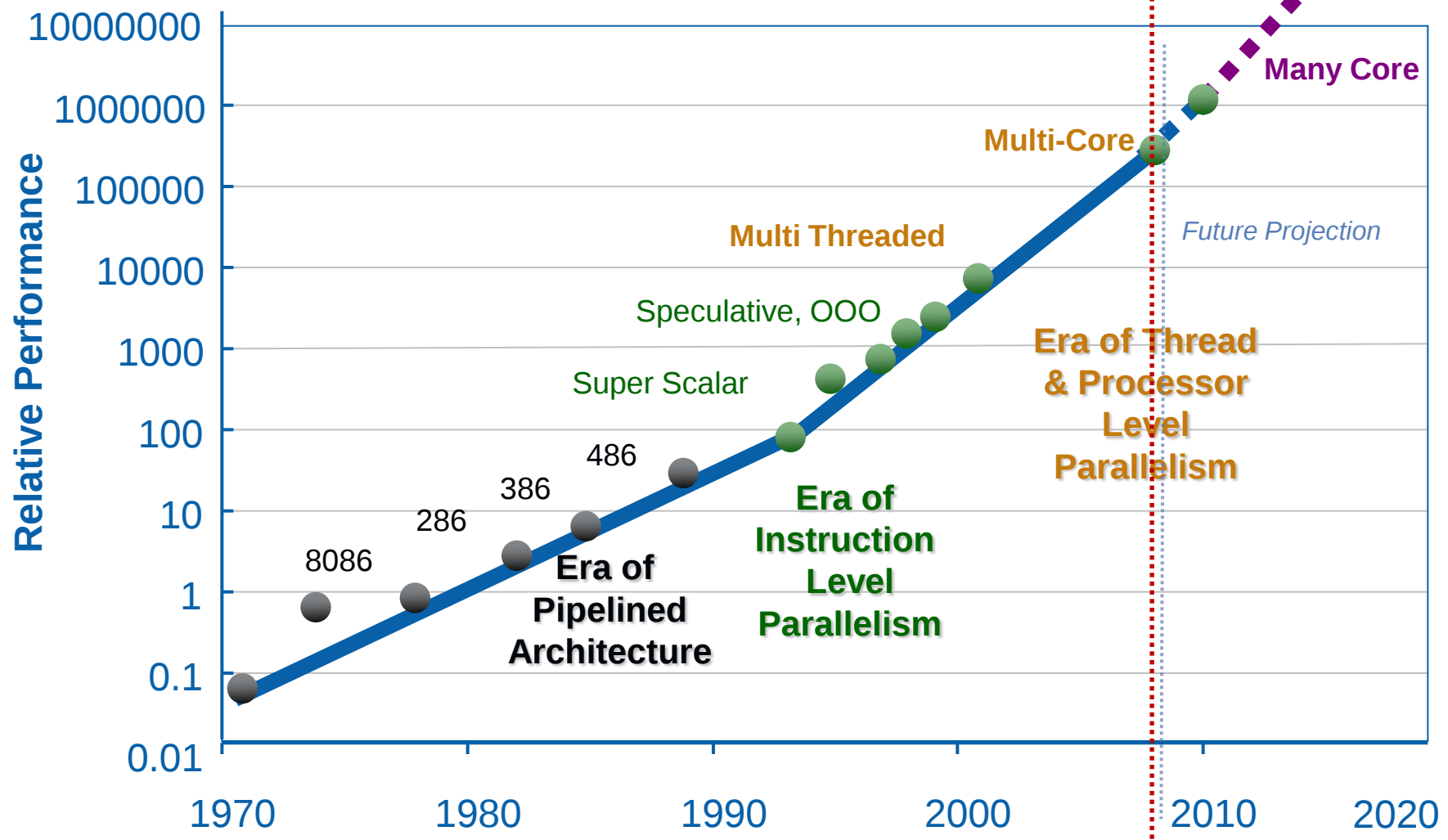
- **Today's Peta: 0.7- 2 nj/op**
- **Today's COTS: 2nj/op**
(assume 100W/50GFlops)
- **Unmanaged Exa: if 1GW, 0.31nj/op;**



Unmanaged growth in power will reach Giga Watt level at Exascale



Parallelism for Energy Efficient Performance



Intel estimates of future trends. Intel estimates are based in part on historical capability of Intel products and projections for capability improvement. Actual capability of Intel products will vary based on actual product configurations.



Специализированные архитектуры

Гибридные платформы



Планы развития систем СКИФ Ряда 4



СКИФ/П

- 200 ТФЛОПС/шкаф
- 1.37 ГФЛОПС/Ватт

СКИФ/С

- 100 ТФЛОПС/шкаф
- 1.0 ГФЛОПС/Ватт

СКИФ/В

- 37 ТФЛОПС/шкаф
- 0.38 ГФЛОПС/Ватт

СКИФ/Н

- 24 ТФЛОПС/шкаф
- 0.25 ГФЛОПС/Ватт

2009

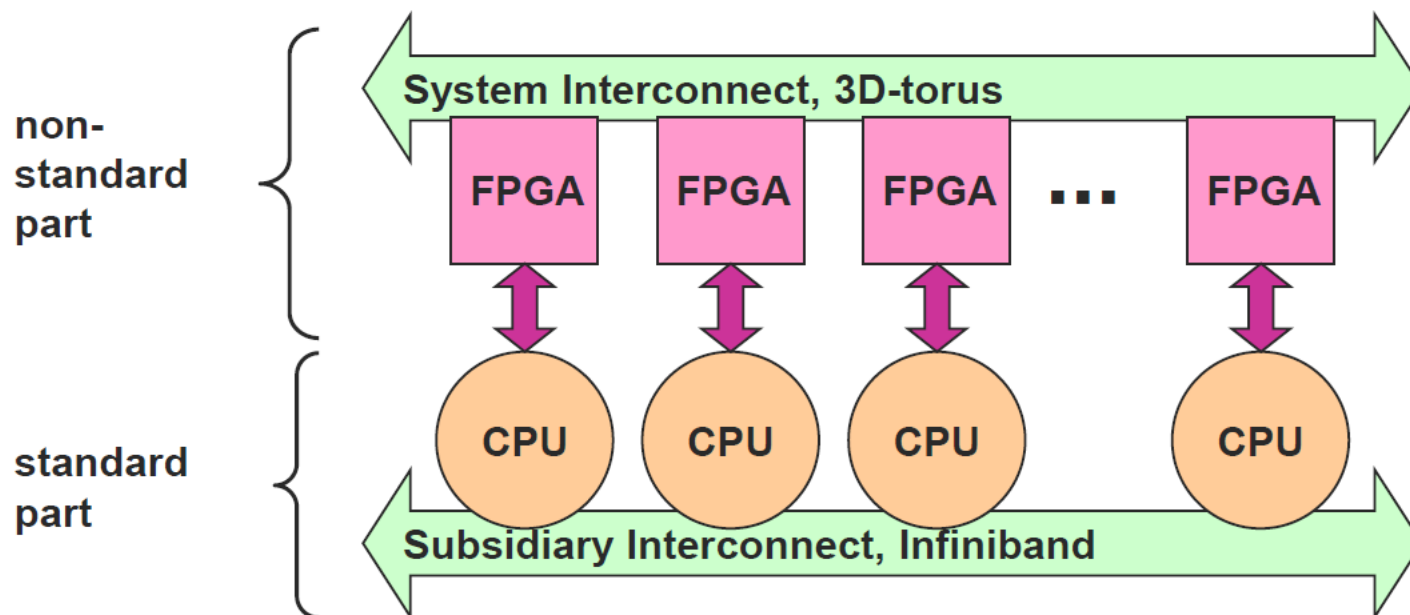
2010

2011



Направления R&D «использование ресурсов FPGA»

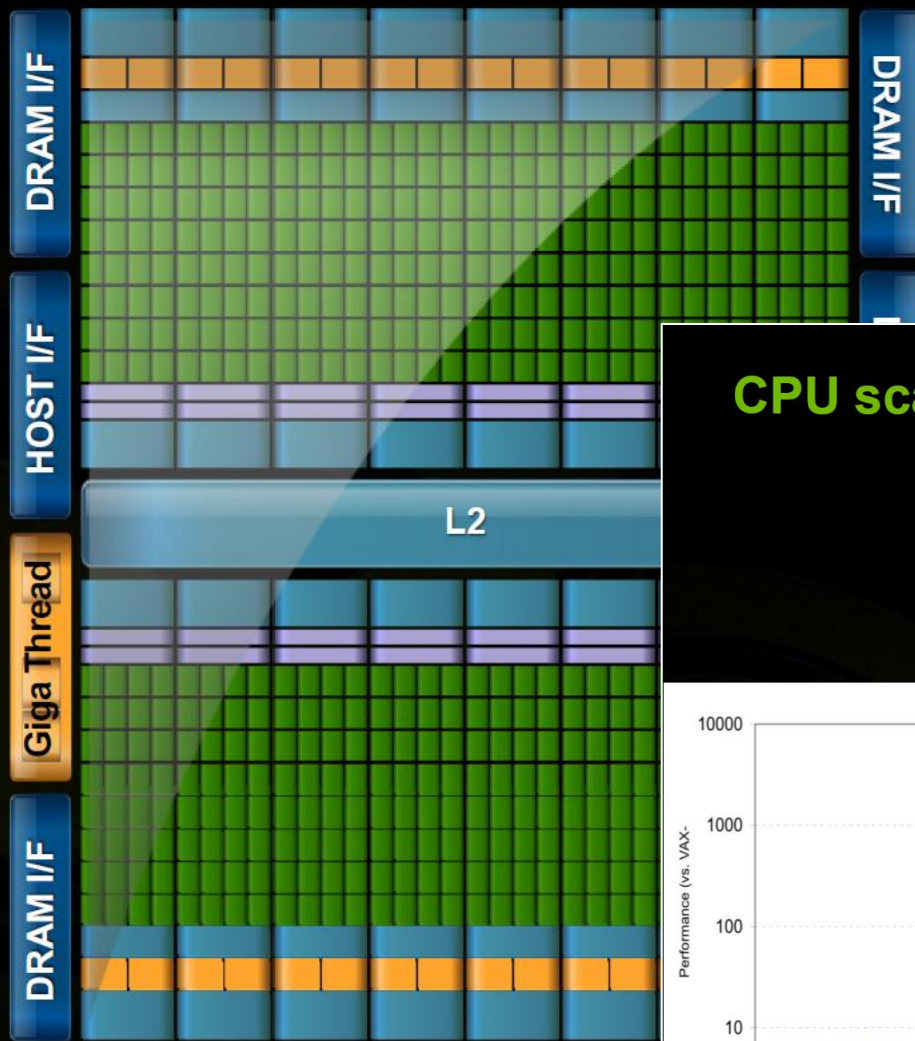
- ★ Реализация коллективных MPI-операций в FPGA
- ★ FPGA-поддержка PGAS-языков (UPC, Titanium, и др.) и других моделей (nccNUMA, односторонние обмены и т.п.)
- ★ Поддержка гибридных (FPGA+CPU) вычислений



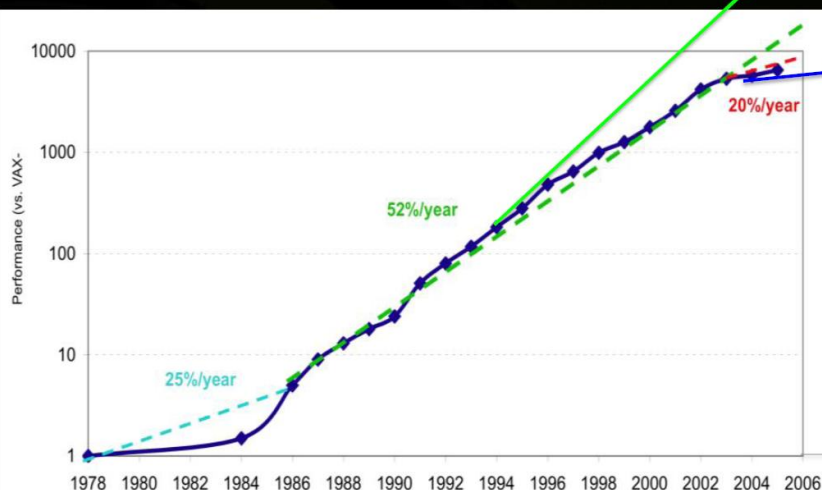
Fermi is a throughput computer



- 512 efficient cores



CPU scaling ends, GPU continues

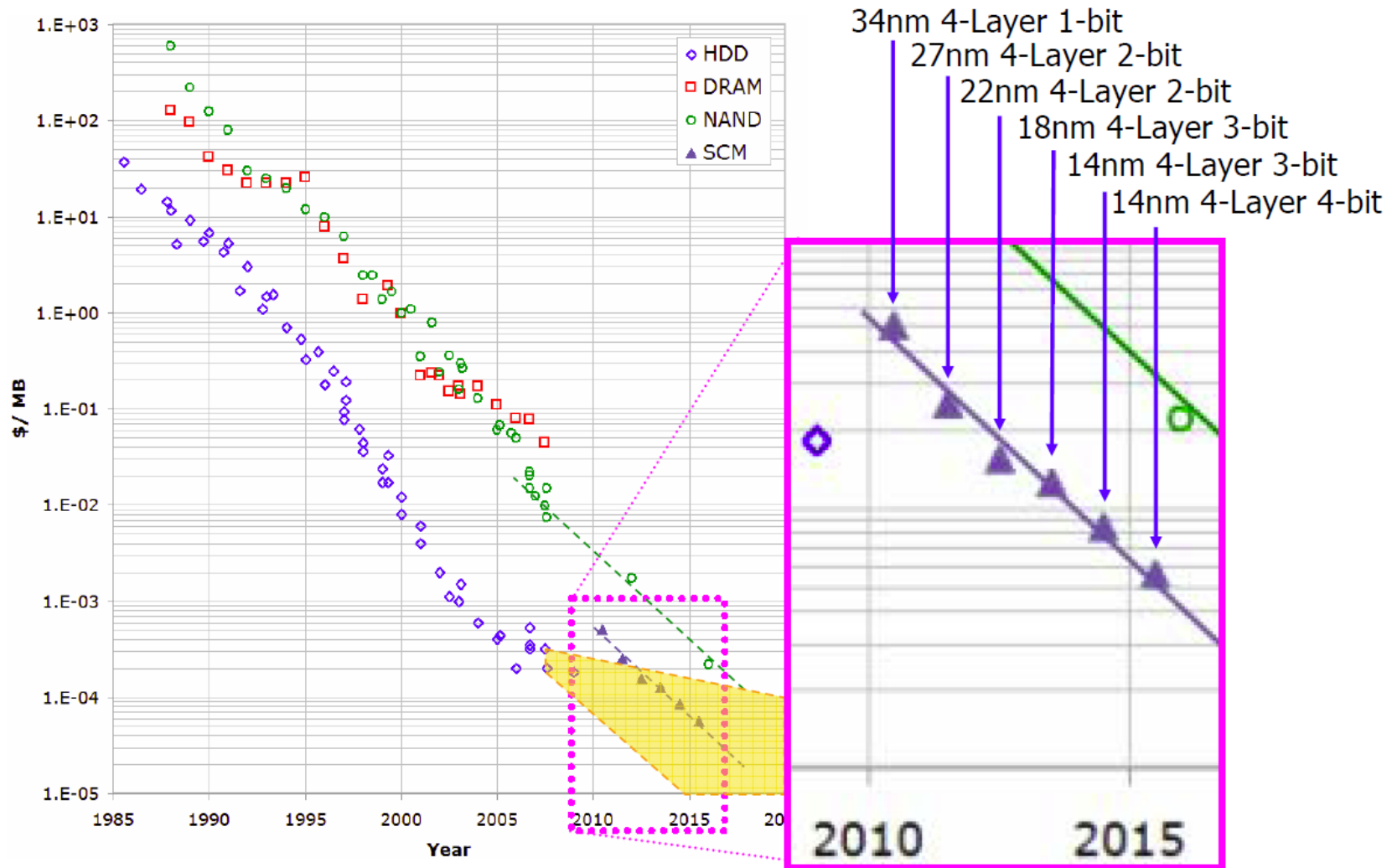


2017

Системы хранения данных

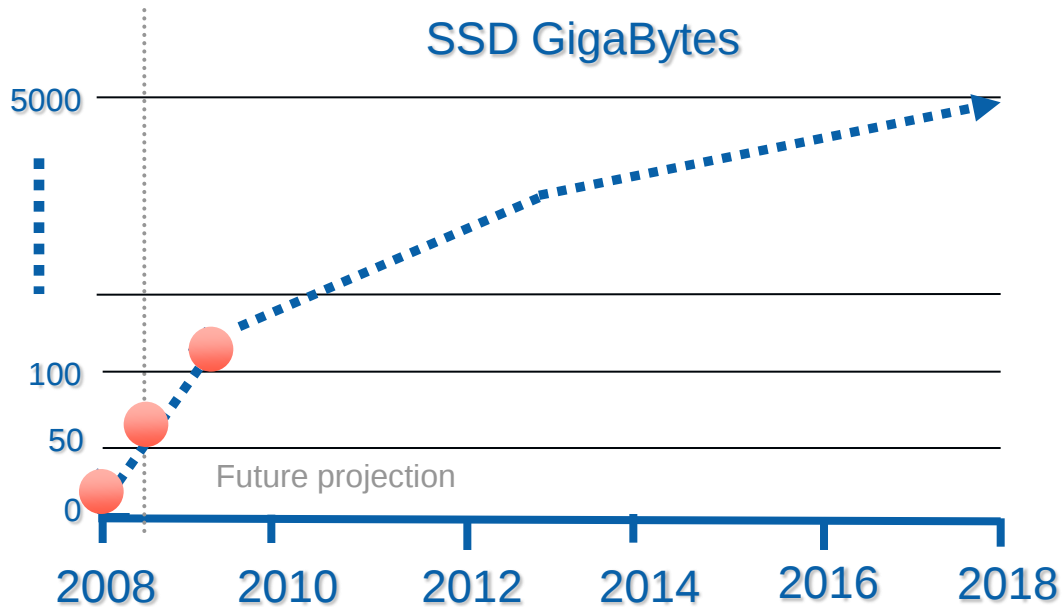
RAM/SSD/HDD/Tape

Storage Historic Price Trend and Forecast



Solid State Drive

Future Performance and Energy Efficiency



Source: Intel, calculations based on today's vision

Assume: Capacity of the SSD grows at a CAGR of about 1.5; historical HDD at 1.6

Vision

10 ExaBytes at 2018:

- 2 Million SSD's vs. ½ Million HDD
- If @2.5w each, total 5MW
- If HDD (300 IOPS) and SSD (10k IOPS) constant: SSD has 140X

Innovations to improve IO: 2X less power with 140x performance gain



Существующее решение: HP ESL-E

5 PB uncompressed / 3500 tapes в 600 мм стойке



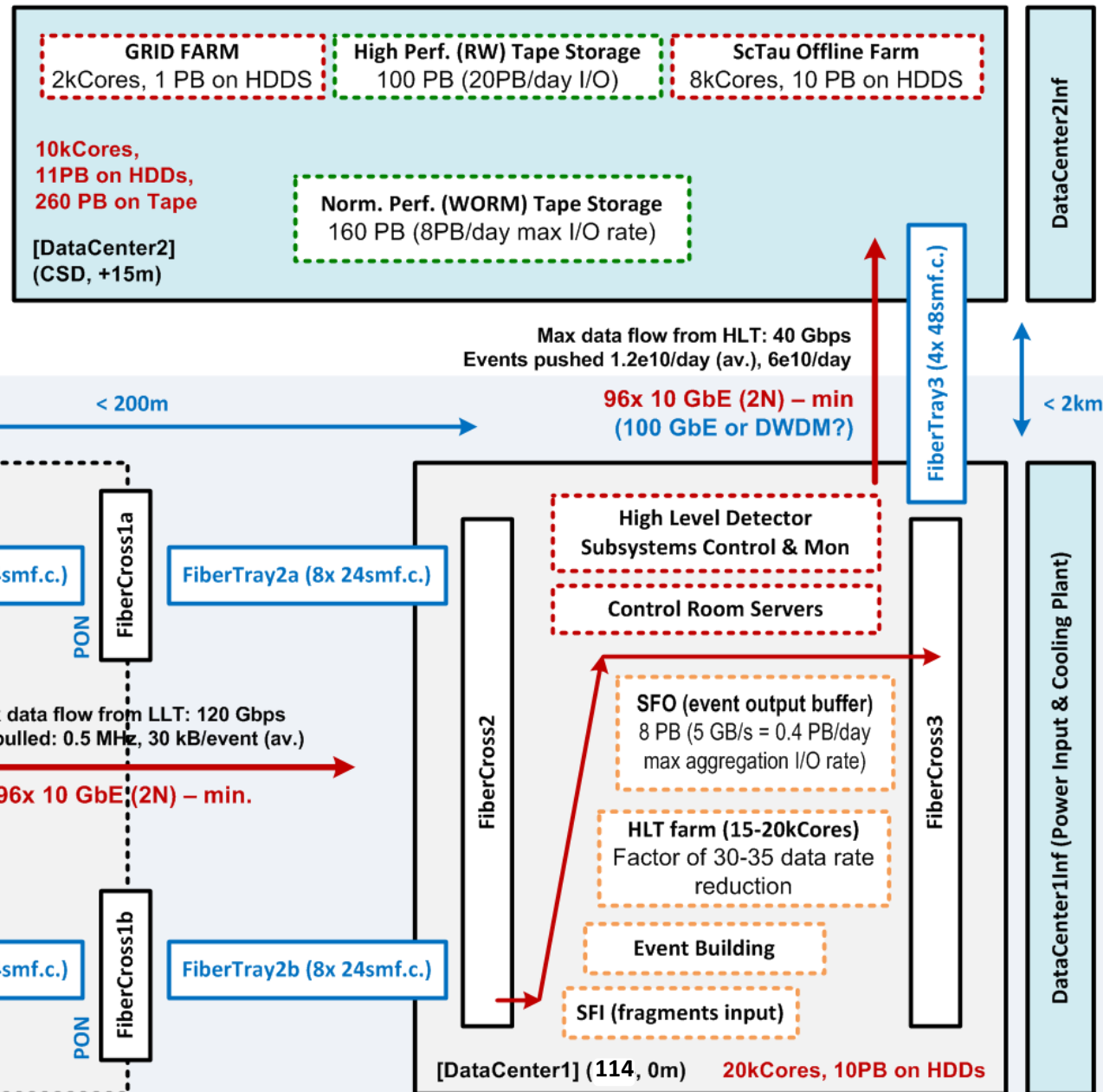
Family Comparison

	ESL 712e	ESL 322e
Announce Date	February 2004	July 2004
Drive Technology	LTO-5 Ultrium 3280 FC, LTO-4 Ultrium 1840 FC, LTO-3 Ultrium 960 FC	LTO-5 Ultrium 3280 FC, LTO-4 Ultrium 1840 FC, LTO-3 Ultrium 960 FC
Maximum Cartridge Count	712 (1 frame) 3546 (5 frame)	322 (1 frame) 3546 (5 frame)
Maximum Drive Count	24 (1 frame) 44 (5 frame)	24 (1 frame) 44 (5 frame)
Maximum Capacity (compressed 2:1)	2136 TB (1frame) 10638 TB (5 frame)	966 TB (1frame) 10638 TB (5 frame)
Maximum Throughput (compression 2:1)	24.2TB/hr (1frame) 44.4TB/hr (5 frame)	24.2TB/hr (1frame) 44.4TB/hr (5 frame)
Interface Manager	Included	Included
ETLA Components: (not included)		

FUTURE: “Oracle unrolls tape roadmap” (11.08.2010)

- “In 2013 we have a fourth generation tape, a T10k-D we'll say, with **6-10TB capacity and 270-400MB/sec throughput**. The SL8500 now has 600-1000PB capacity and 621-920TB/hr throughput. Once again the slot count stays constant and so does the number of drives, using the same math as before.”
- “The gen 5 tape, a T10k-E, appears in 2015 with a **12-20TB capacity and 400-600MB/sec throughput**. At this point the SL8500 has a **1.2EB (1,200PB)-2EB (2,000PB) capacity and a 920-1380TB/hr throughput**, the slot and drive count remaining constant again.”

Общие характеристики
компьютерного обеспечения
системы сбора данных детектора
для супер с-tau фабрики



IT обеспечение (за исключением оборудования на антресолях детектора)

- 2 физически независимые площадки, разнесенные на 500 м (машзал в здании детектора + машзал ОВС)
 - 2N резервирование системы питания и системы управления, N+1 – резервирование системы охлаждения на обеих площадках
 - Группа специализированных систем ИБП на каждой площадке (использование дизель-генераторов не предполагается)
 - Высокая степень автоматизации инженерной инфраструктуры
- **Полезная нагрузка:**
 - 30k ядер CPU (0.6 PFlops, 20 TFlops/CPU core в 2020 г.):
20k в помещении 114 + 10k в ОВС (включая ГРИД-ферму)
 - 20PB на “дисковых системах” (HDD/SSD, буфер для передачи данных в помещении 114 (10PB) + буфер offline-фермы детектора в ОВС (10PB)
 - 80PB на “быстром ленточном хранилище”: 20 ленточных роботов в ОВС
 - 160PB на “медленном ленточном хранилище”: 32 ленточных робота в ОВС

IT обеспечение (за исключением оборудования на антресолях детектора)

- 20PB (HDDs) + 80PB (Tapes) + 30k CPU cores позволяют:
 - Сохранять для долговременной обработки в среднем около 3% всех экспериментальных событий, отобранных первичным триггером детектора при проектной максимальной возможной частоте его срабатывания 0.5 MHz, что соответствует примерно 3×10^{12} сохраненных для длительной обработки экспериментальных событий за 5 лет непрерывной работы детектора (10 лет астрономического времени при 50% эффективности)
 - Иметь возможность, при необходимости, сохранять для дальнейшей обработки весь поток данных, идущих с первичного триггера (120 Gbps)
 - Иметь на одно экспериментальное событие, сохраненное в системе хранения данных offline фермы 1-3 kFlops вычислительной мощности.
- Необходимый бюджет на 5 лет непрерывной работы:
 - Вычислительные системы и системы хранения данных: 13 M\$
 - Эксплуатационные расходы (в том числе ремонт IT оборудования и его обслуживание, расходы на электроэнергию): 2M\$
 - Стоимость инженерной инфраструктуры: 4-5 M\$

Для дальнейшей детализации и обоснования предлагаемого IT решения необходимо: (1)

- **Выбрать архитектуру системы компоновки события из фрагментов (Event Builder-a):**
 - Стандартные средства пакетной коммутации на базе IP/Ethernet (**высокая гибкость, высокая стоимость, пониженная производительность**)
 - Системы на основе специализированных коммутаторов/сетевых протоколов FC, FCoE, FCoIP (**высокая производительность, высокая гибкость, высокая стоимость**)
 - Специализированные решения на основе многопроцессорных систем с общей памятью/специализированного ПО (**наивысшая производительность, сравнительно низкая стоимость оборудования, необходимость разработки и поддержки ПО**)
- **Выбрать архитектуру третичного триггера и фермы offline-обработки данных:**
 - Архитектура общего назначения (**простота разработки ПО, высокая переносимость, сравнительно низкая эффективность**)
 - Специализированная гибридная архитектура (**сложность разработки ПО, низкая переносимость, высокая эффективность**)

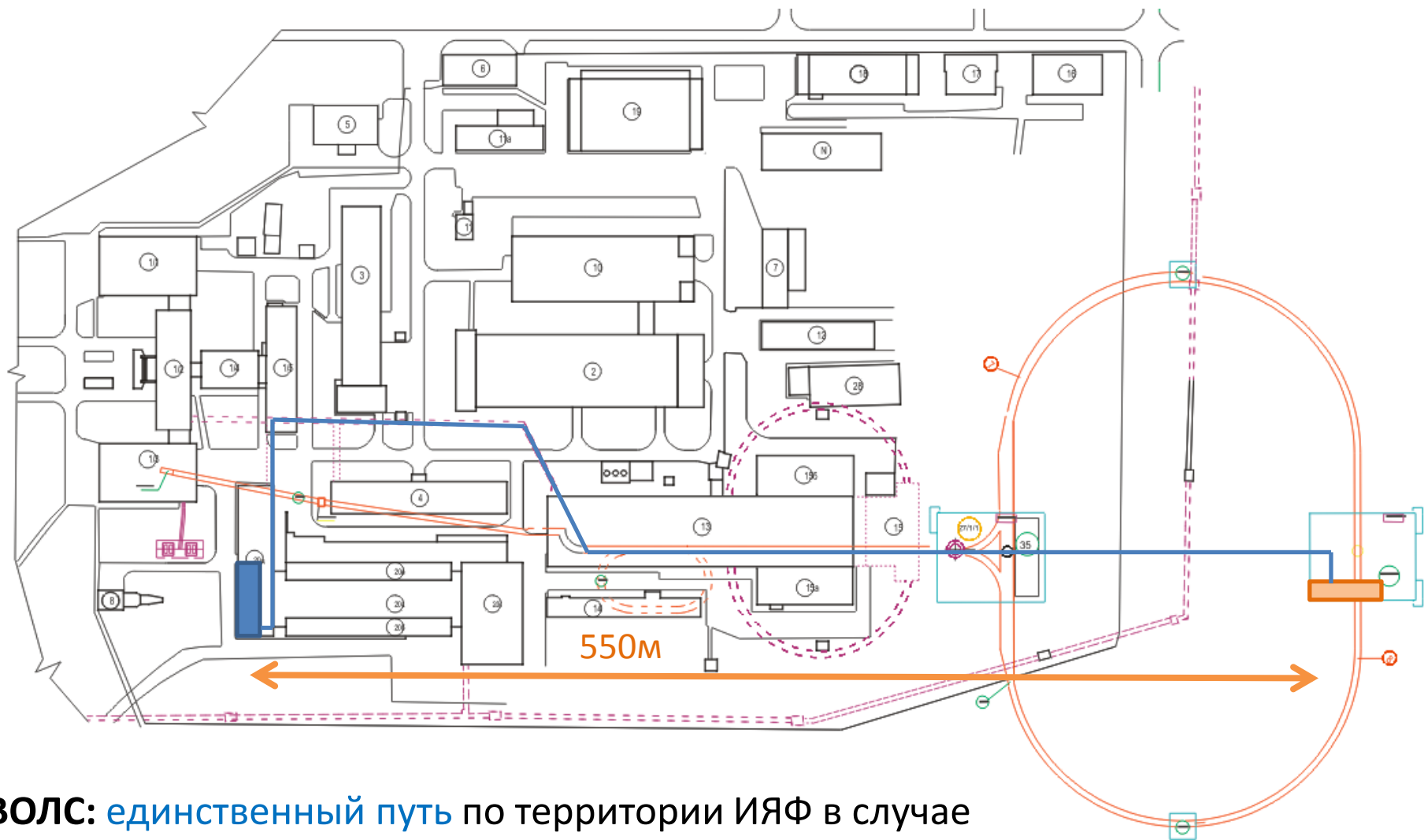
Для дальнейшей детализации и обоснования предлагаемого IT решения необходимо: (2)

- **Определиться с архитектурой ПО EventBuilder-a, третичного триггера, реконструкции и моделирования событий детектора:**
 - Однопоточное/многопоточное ПО?
 - Требования к внутренним сетям? (пропускная способность, задержки, необходимость высокоточной синхронизации)
 - Ограничение снизу на объем RAM на единицу вычислительной мощности?
- **Определиться со стандартной средой исполнения ПО детектора:**
 - Интенсивное использование виртуализации со сравнительно небольшим числом одновременно доступных ядер CPU (упрощение процесса разработки, высокая гибкость и переносимость решения, несколько пониженная производительность)
 - *либо* ориентированность на среды исполнения с одновременной доступностью большого числа ядер CPU (много больше 32-х) (высокая производительность, ограниченность в выборе ОС, усложнение процесса разработки, низкая переносимость)

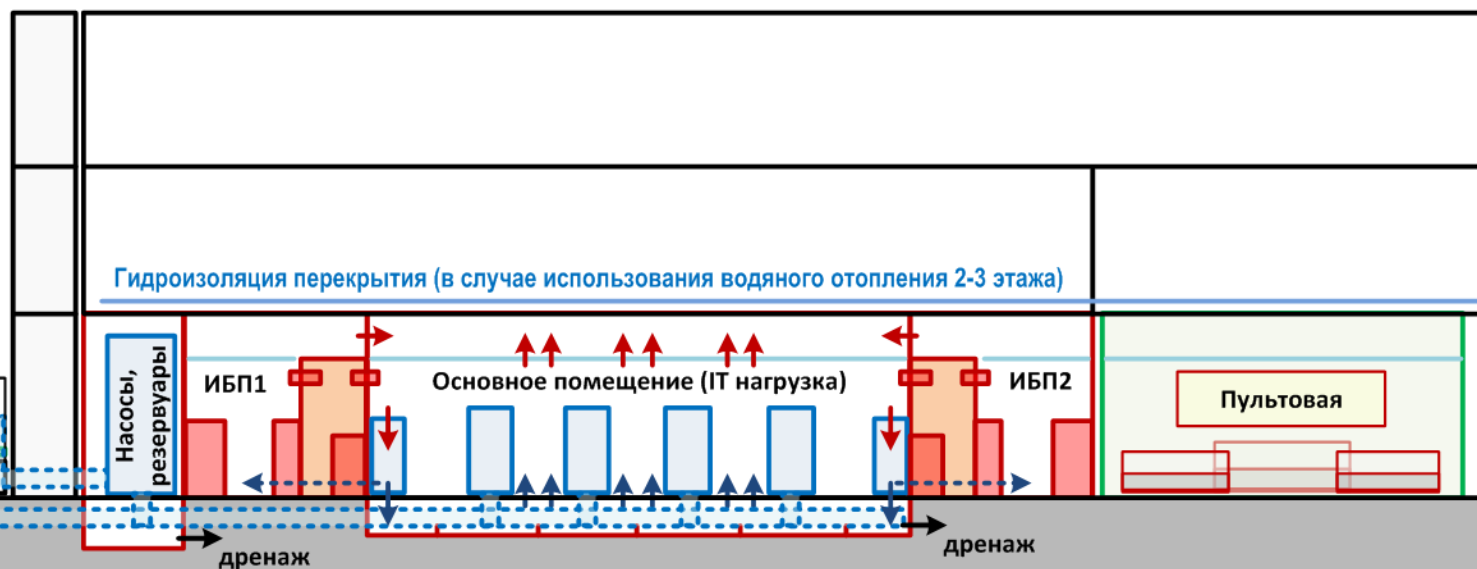
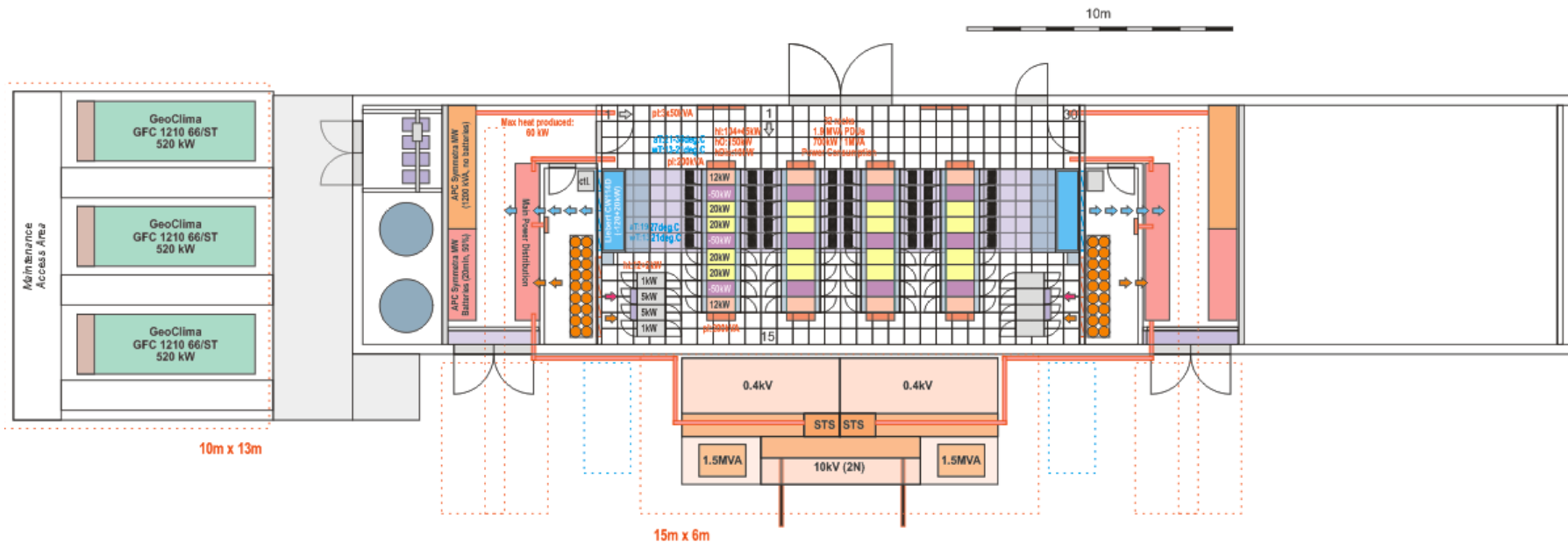
Для дальнейшей детализации и обоснования предлагаемого IT решения необходимо: (3)

- **Определиться с высокоуровневой архитектурой системы хранения и offline-обработки данных:**
 - GRID система? (высокая масштабируемость и отказоустойчивость, сложность развертывания и администрирования, необходимость в наличии внешних каналов связи)
 - Консолидированный центр обработки данных (отсутствие дорогостоящих каналов связи, высокая управляемость и простота администрирования, ограниченная масштабируемость и сравнительно низкая отказоустойчивость)?
 - Консолидированный центр обработки данных с поддержкой совместимости с распространенными Cloud-платформами (использование свободной емкости региональных суперкомпьютерных центров)?

Объекты ИТ инфраструктуры детектора



ВОЛС: единственный путь по территории ИЯФ в случае использования LR 10-100 GbE/SMF fiber или до 4-х независимых путей в случае использования технологии DWDM



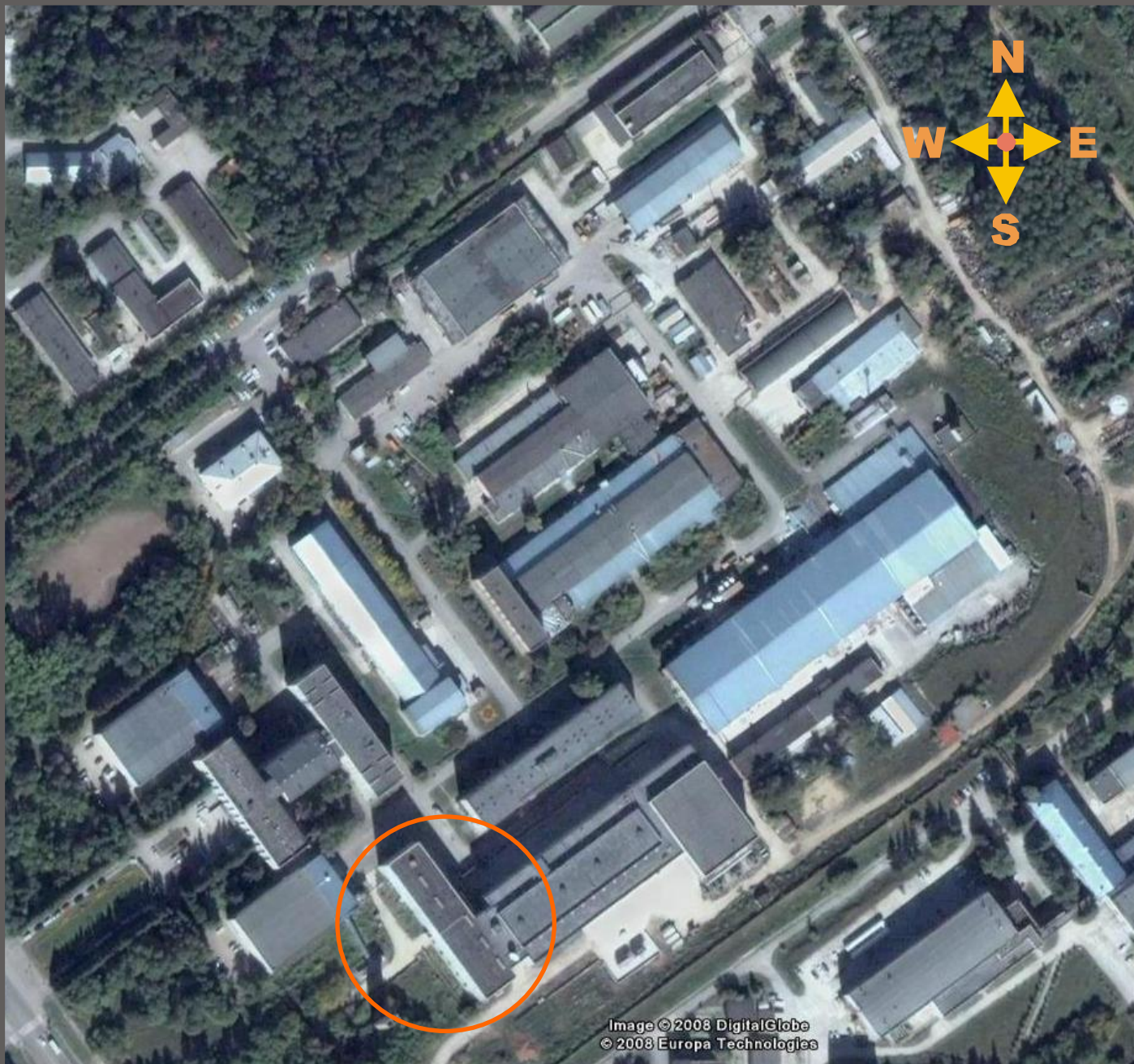


Image © 2008 DigitalGlobe
© 2008 Europa Technologies



- Legacy Free Standing Servers
- Legacy Open Frame Rack
- Legacy Compartment
- APC Open Frame Rack (43U)
- HP ESL-E Tape Library (42U)
- Floor Plates To Be Replaced

False Floor Plate Counters

Full size plates (0.36m²): 973
 Half size plates (0.18m²): 65
 1/4 size plates (0.10m²): 24
 1/8 size plates (0.05m²): 37

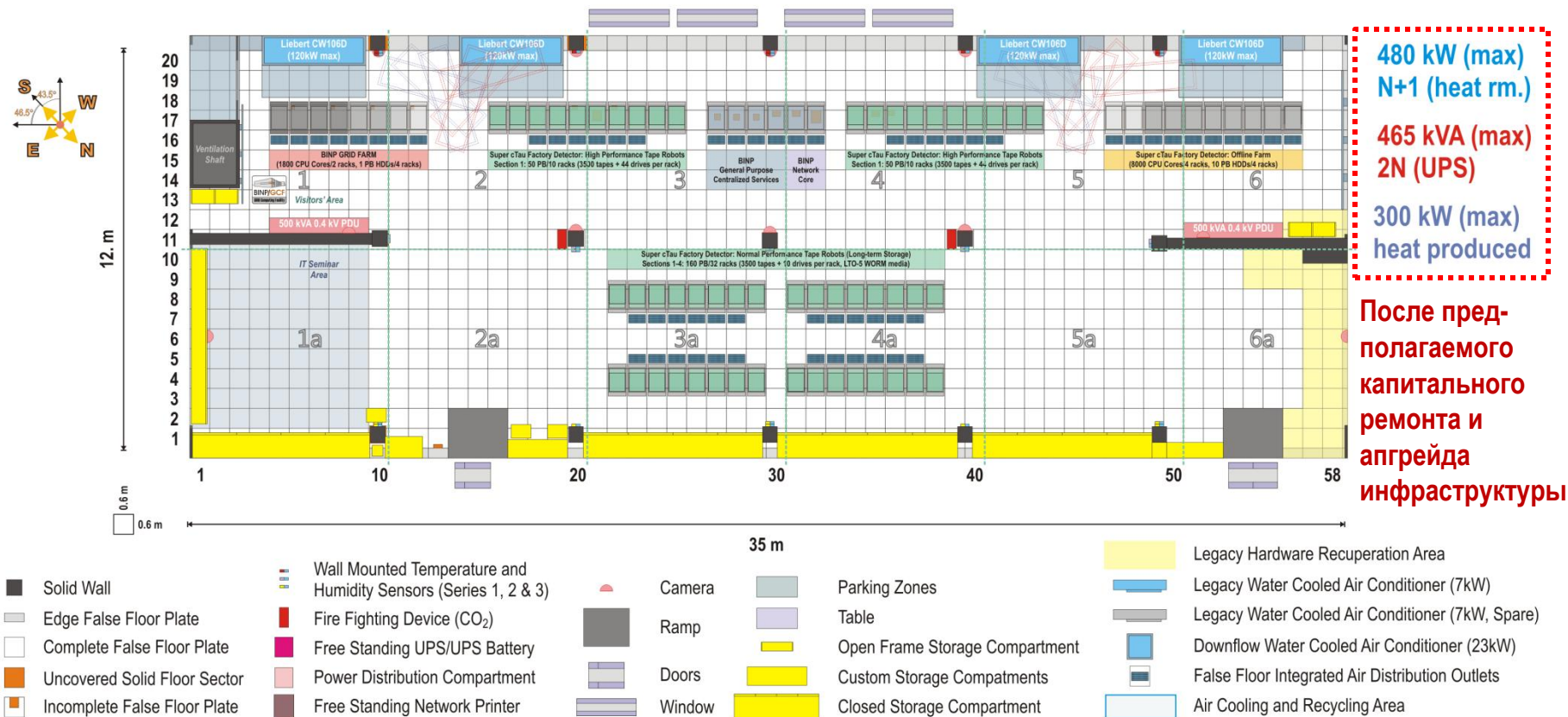
Area Estimates

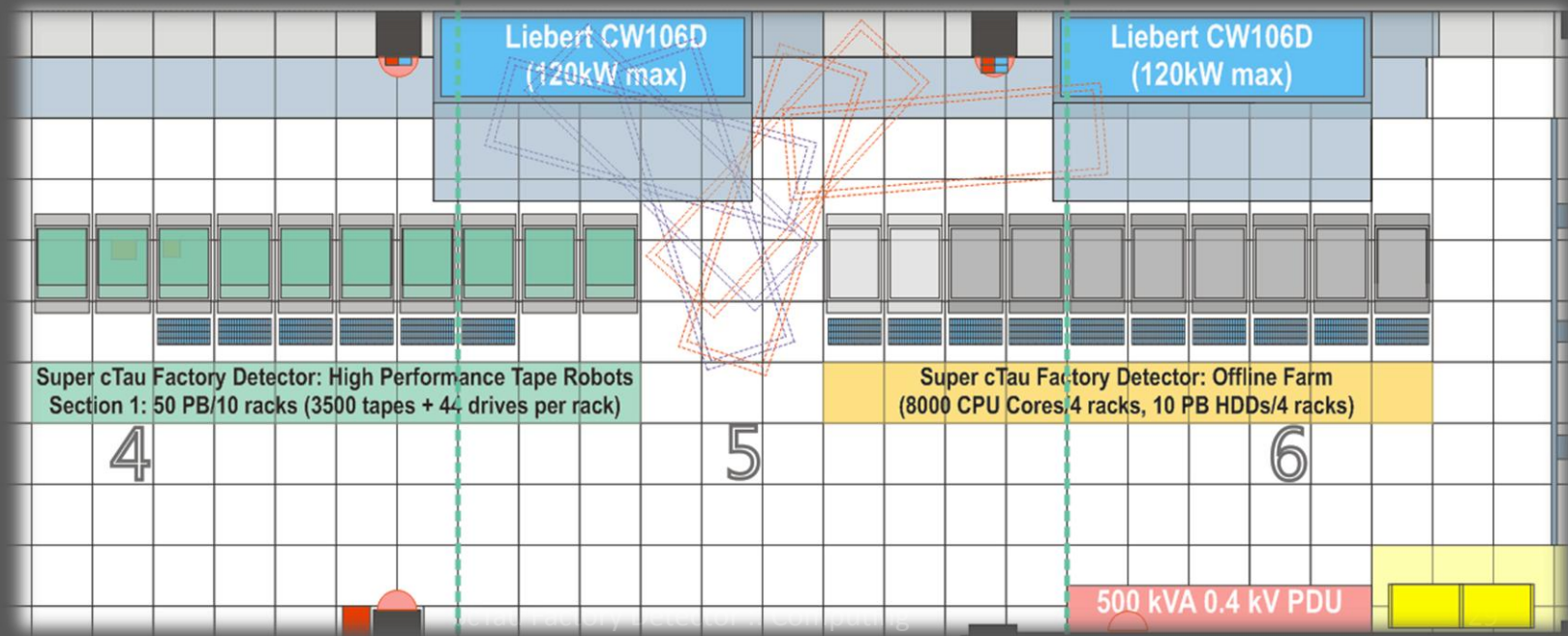
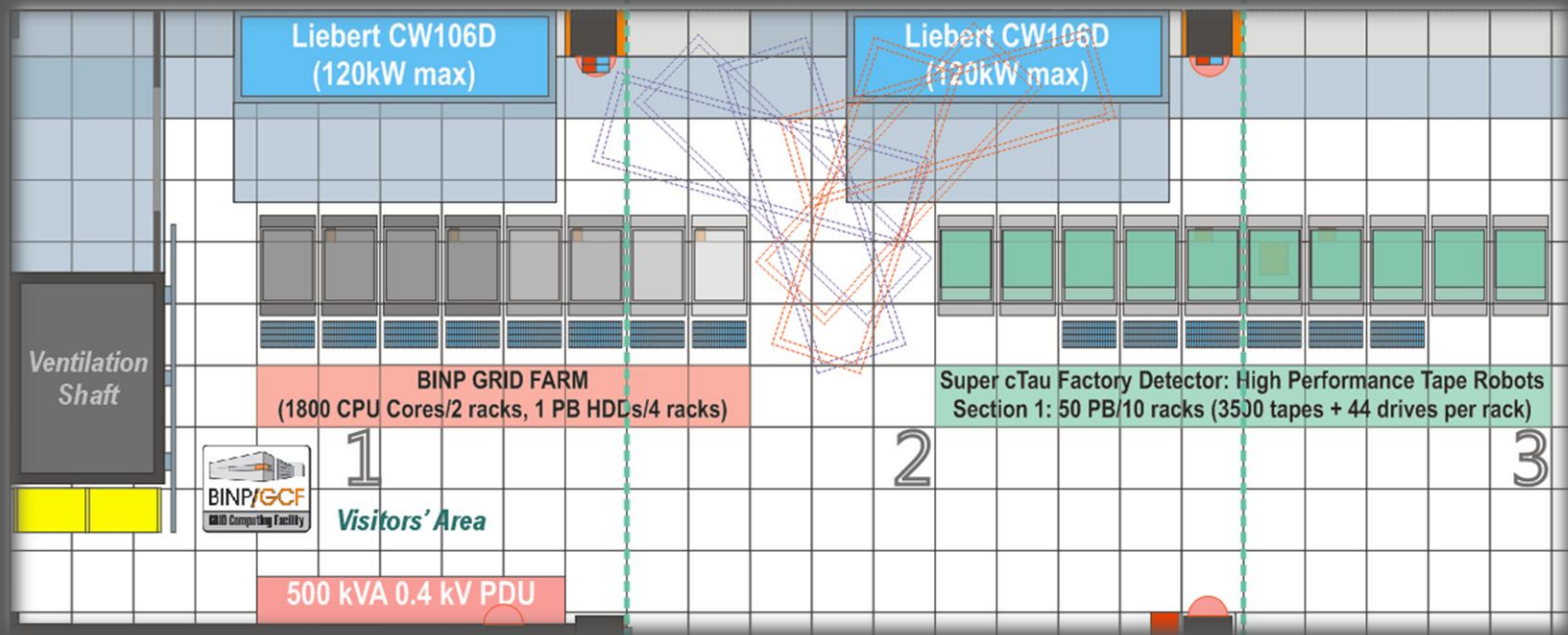
Total usable area: 366.2m² (of 440m² within the outline)
 Legacy hardware zone area: 85.4m² (23%)

BINP IT FACILITY MAP

Alexander S. Zaytsev <A.S.Zaytsev@inp.nsk.su>

2020



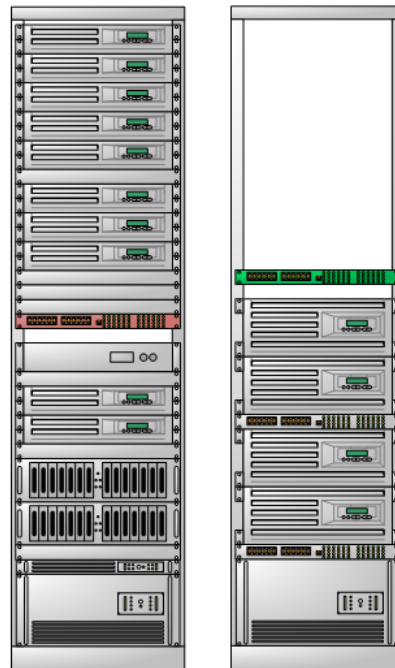


Предлагаемые первые шаги реализации и обоснованию предложенных решений

Создание прототипов и отдельных
компонентов системы (2010 – 2014):
три основных направления

Поэтапное наращивание вычислительной и телекоммуникационной инфраструктуры

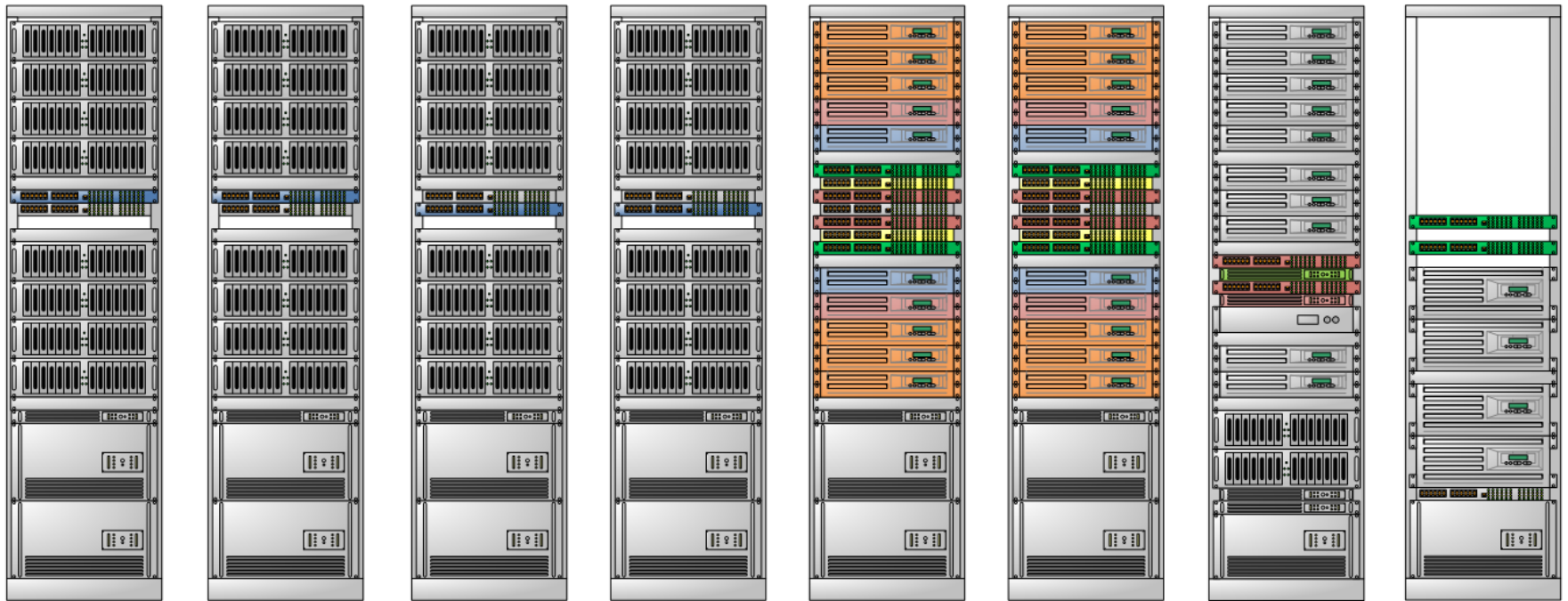
- ГРИД-ферма как прототип фермы offline-обработки данных детектора и полигон для отработки технологий построения систем хранения данных:
 - 10-100 GbE, оптимизация производительности сетей
 - Коммутируемые сети хранения данных (SAN/FC, FCoE)
 - Параллельные файловые системы (PVFS2, Lustre и прочие гетерогенные системы с выгрузкой метаданных на отдельные высокопроизводительные группы носителей)
 - Масштабное использование виртуализации, в том числе и с поддержкой “прозрачной” миграции VM-ов, между (географически) удаленными вычислительными комплексами
 - Получение опыта эксплуатации системы масштаба **2k CPU cores/1 PB HDD** (3-5% от финальной конфигурации)
 - Получение опыта эксплуатации системы ИБП/распределения питания до **150 kVA** (4-5% от конечной конфигурации)



40 Cores (100 kSI2k) / 200 GB RAM / 28 TB (23 TB visible)

3.4 MRub
Accumulated
since 2004

2010Q3



1832 CPU Cores (5.5 MSI2k) / 4.8 TB RAM / 1.0 PB HDD

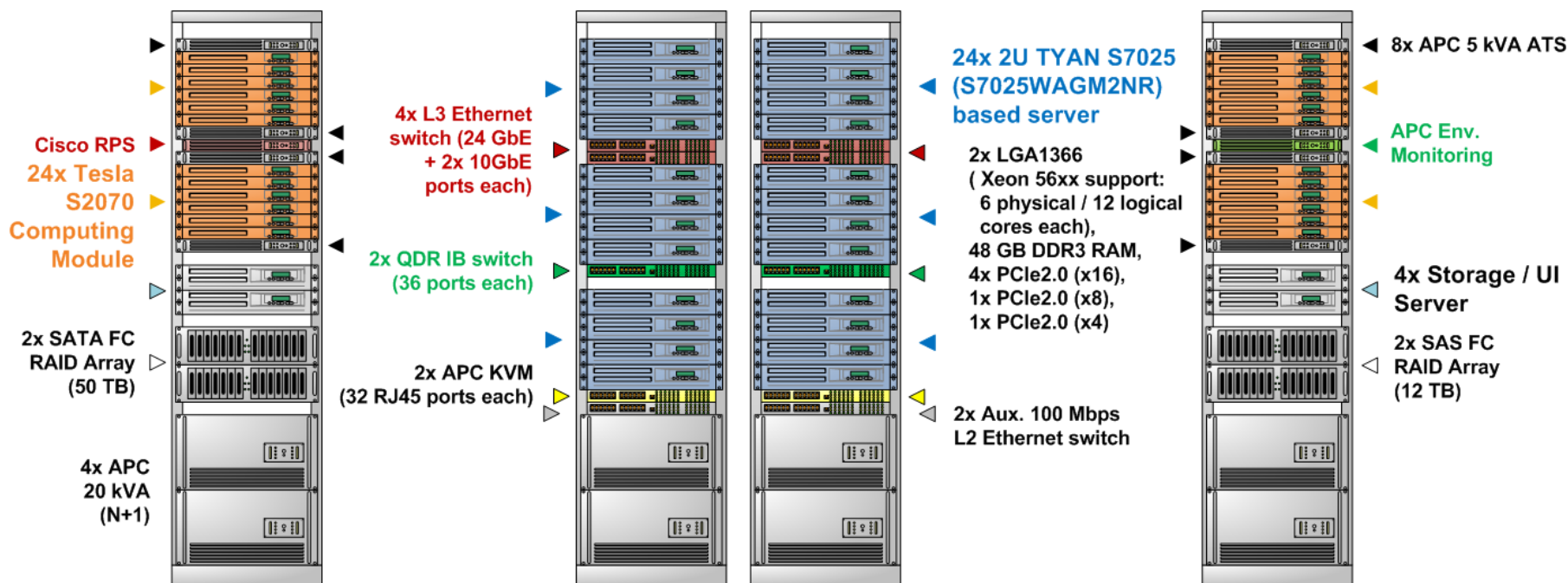
3.2 M\$
(w/o PDU & cooling
infrastructure)

2012-2013

Поэтапное наращивание вычислительной и телекоммуникационной инфраструктуры

- Прототипирование решений для третичного триггера и EventBuilder-a на основе гибридных архитектур (CUDA/NVIDIA Tesla) – *в случае выбора соответствующей архитектуры в качестве основной*
 - Удельная стоимость в 5 раз ниже решения на основе платформ общего назначения (по состоянию на 2010 г.)
 - Доступность средств разработки и отладки ПО
 - Полноценная поддержка DP (начиная с NVIDIA Fermi, 2010 г.)
 - Высокая энергоэффективность и компактность решения
 - Возможность построения первой очереди решения при сравнительно небольших затратах (10 млн.р.)
 - Возможность использования топологий 2D и 3D-тор для системной сети
 - *Основные недостатки данного подхода были перечислены выше*

Annotated rack layout: front view
(no blanking panels and cable organizers shown here)



Specs summary:

24 x 2 x 12 = **576** logical CPU cores (Xeon W5680, 3.3 GHz, 3.8 TFlops)
 24 x 4 x 448 = **43008** CUDA cores
 24 x 24 GB = **0.56 TB** of GDDR5 VRAM
 24 x 48 GB = **1.13 TB** of DDR3 RAM
 24 x 80 GB = **1.88 TB** on SSD drives (80 mks latency for reading)
 4 x 5 x 146 GB = **2.9 TB** on 15000rpm SAS drives in RAID6
 24 x 146 GB = **3.4 TB** on 15000rpm SAS drives in RAID10
 12 TB on 10000rpm SAS drives in RAID6
 50 TB on 7200rpm SATA drives in RAID6
 [70 TB of storage in total]

24 x 4 Tesla C2070 GPUs: **53 - 66 TFlops** (double precision)
 107 TFlops (single precision)

Hardware overall cost estimate:

Total: **1.23 M\$** (35.7 MRub)

\$ 19 - 23 / (1 Gflops + 1.2 TB)
 for GPU based design (double precision)

Compare to **\$ 100 - 110 / GFlops** (double precision)
 for the general purpose CPU based design

Front view: *plain layout*



A

B

C

D

Rack (UPS) IDs

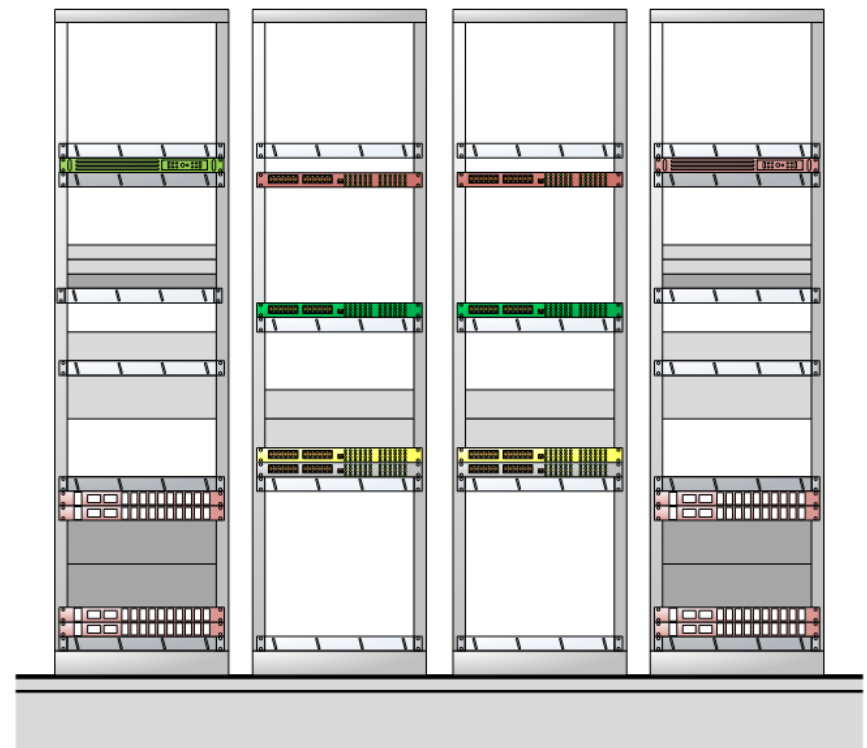
4 x 4 Tesla C2070 GPUs: **8.8 - 11 TFlops** (double precision)
18 TFlops (single precision)

[35 TB of storage in total]

Hardware overall cost estimate: 349 k\$ (10.1 MRub)

Open frame rack based layout
(step 1: building 17% testbed)

Rear view: *data and power cable organizers*
Planar (switched fabric based) layout



D

C

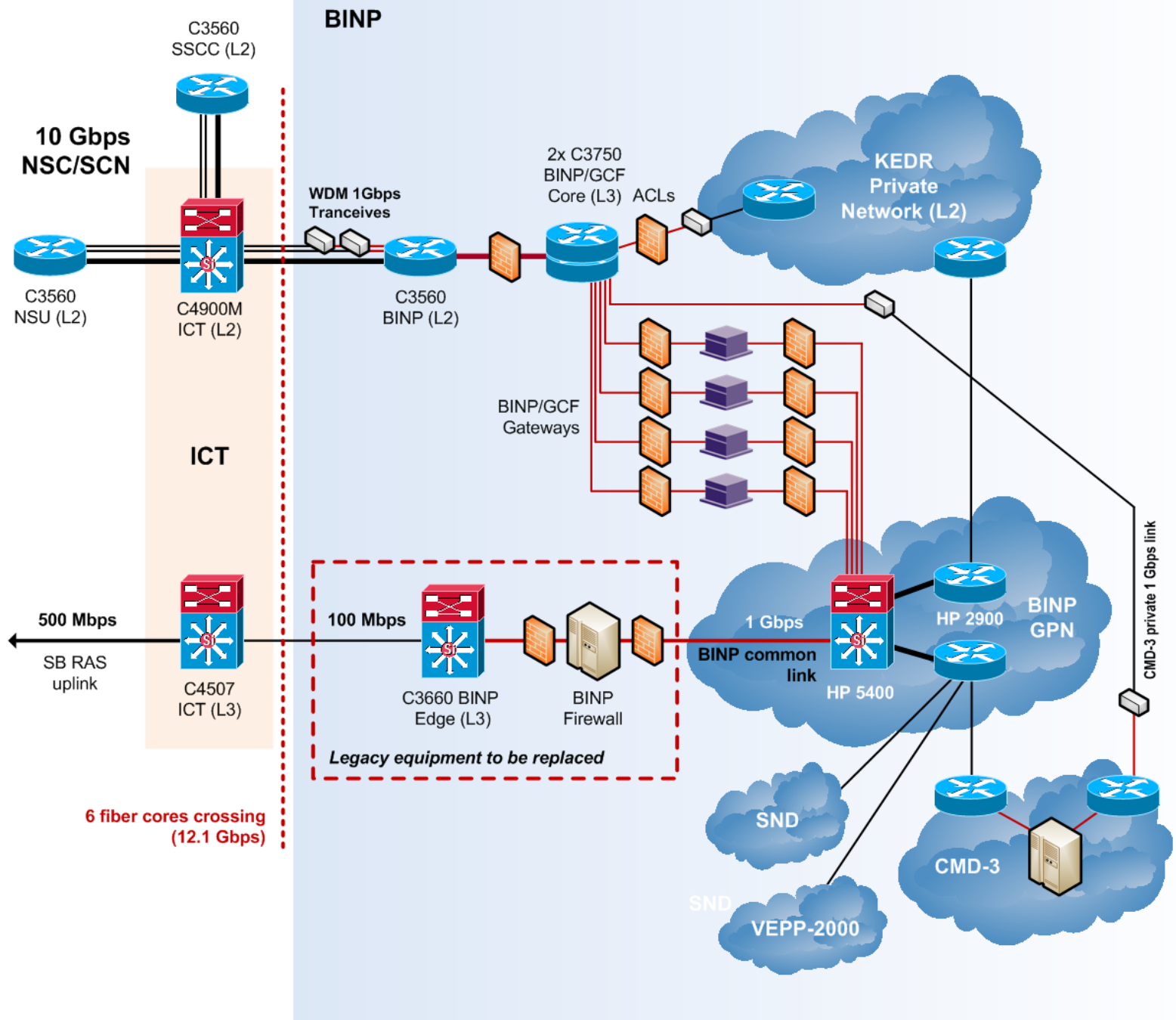
B

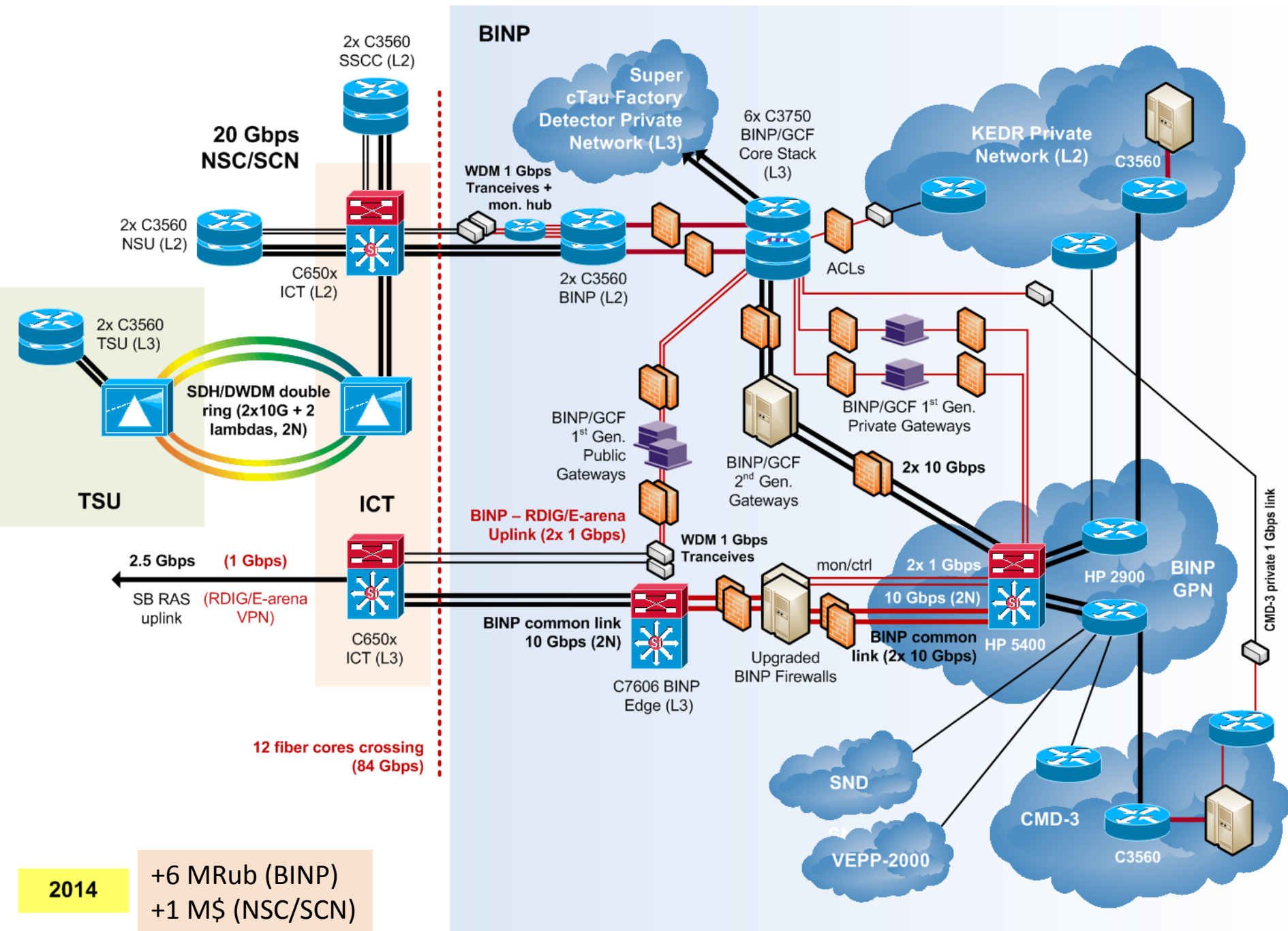
A

Rack (UPS) IDs

Поэтапное наращивание вычислительной и телекоммуникационной инфраструктуры

- Модернизация телекоммуникационной инфраструктуры ИЯФ и ГРИД-фермы ИЯФ:
 - выход во внешний мир для всей IT инфраструктуры детектора
 - привлечение географически распределенных вычислительных ресурсов
- Заблаговременная модернизация машзала ОВС:
 - 2x 500kVA – распределение питания
 - 480kW – отводимая тепловая мощность
 - контроль доступа и системы мониторинга
- Заблаговременная отработка решений для систем управления и мониторинга машзалов для будущего детектора (системы дискретных и распределенных датчиков):
 - Сторонние решения (APC)
 - Собственные решения (ИЯФ)
 - *Возможный совместный проект ИАЭ по созданию собственного решения для распределенного рамановского оптоволоконного термометра*





QUESTIONS & DISCUSSION