Аппаратные решения для обработки Больших Данных в большой памяти. HP VMA

Кирилл Вахрамеев,
специалист по продвижению серверных решений для критически важных приложений,
OpenVMS Ambassador, HP Россия

Аппаратные решения для обработки Больших Данных в большой памяти. HP VMA

Системы повышенной надежности на платформе x86 для аналитики:

- HP ProLiant DL980 + HP VMA

Кирилл Вахрамеев,
специалист по продвижению серверных решений для критически важных приложений,
OpenVMS Ambassador, HP Россия

О чем мы говорили в сессии про Vertica сегодня?

- **Революционная платформа для аналитики в реальном времени** – спроектирована для решения задач завтрашнего дня, доступна уже сегодня
- **Проста в использовании** – быстрая отдача для бизнес-пользователей, DBAs, и программистов
- ***Универсальная СУБД НЕ подходит для Больших Данных*** – *система должна быть специализирована и интегрирована*
- **Производительность, гибкость** – ключевые факторы (и «кубы» строить не нужно)

Большие Данные это сколько?

Размер и классификация хранилища данных, сегодня

- <500ГБ – Маленькое
- 500ГБ > 20ТБ – Типовое
- 20ТБ > 50ТБ – Большое
- >50ТБ – очень Большое
- Несколько лет назад хранилище размером больше нескольких ТБ было редкостью

Большие Данные это сколько и чего?

На самом деле, размер – только один из факторов.

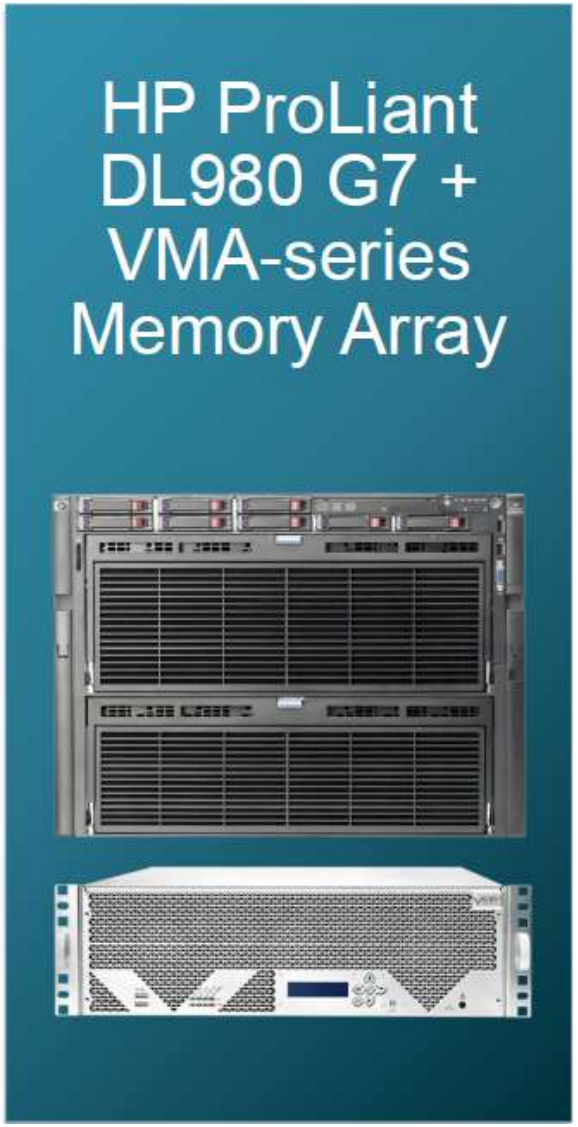
Мир данных трехмерен:

- размер (не помещается)
- скорость (не прокачивается)
- многообразие (не формализуется)

--- Большие Данные – это такие данные, что вышеуказанные факторы не позволяют обрабатывать их «обычными» методами

А если вся программная аналитика уже есть **сегодня, а данные стали Большие?**

- **Нужна аппаратура** – спроектированная для решения задач завтрашнего дня, доступная уже сегодня
- **Простая в использовании и внедрении** – быстрая отдача для бизнес-пользователей, DBAs, и программистов
- ***Универсальная СУБД - это данность***
- **Производительность, гибкость** – ключевые факторы: Большая Система?
- (а «кубы» строить таки придется?)



Масштабируемость, Надежность и Живучесть:

Преодолей парадигму

«Жизнь трудна, по счастью – коротка»!

HP ProLiant DL980G7 – флагман мира x86

PREMA Architecture

Сбалансированная масштабируемость

- Рекорды производительности благодаря **Smart CPU Caching**

Самовосстанавливающаяся структура

- **Отказоустойчивая коммутируемая системная сеть в 2 раза надежнее, чем прямые связи**

Опыт ProLiant

Непревзойденная эффективность

- Общая инфраструктура мира ProLiant: **iLO3, Insight Control, Thermal Logic**

«стандартный» x86 для критически важных задач

Задачи для HP ProLiant DL980 G7

1. Большие базы данных - OLTP и OLAP
2. Консолидация и виртуализация
3. Приложения, требующие особо интенсивного использования процессоров, памяти и ввода-вывода
4. Платформа для критически важных систем

Сервер HP ProLiant DL980 G7

HP PREMA: архитектура для уверенного масштабирования

PREMA архитектура

Отказоустойчивость через самовосстановление

Надежность отдельного сервера 200%, по отношению к «обычному» Proliant

Резервные системные коммутаторы для безотказной работы

Сбалансированная конфигурация

Лидирующая производительность HP Smart CPU caching

До 8 Intel® Xeon® E7 (QPI до 6.4GT/s)

До 10 ядер и 30MB L3-кэш в каждом

До 128 DDR3 DIMM слотов; **до 4TB памяти!**

До 16 слотов, PCI-E 2.0 или PCI-X

4-х встроеных порта 1GbE (опционально 2x 10GbE)

Невероятная эффективность

Консолидируйте до 300 машин на одной системе

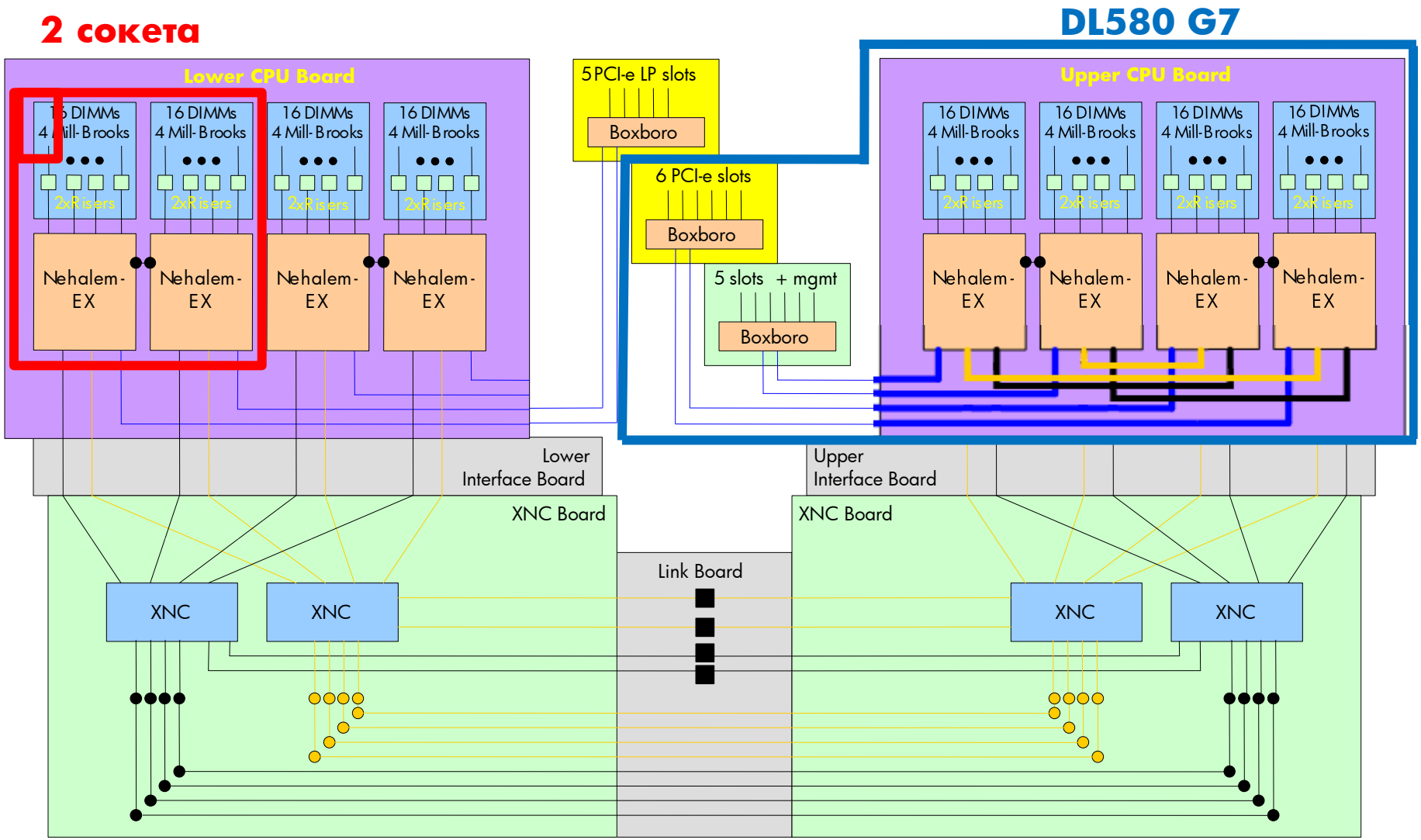
Инновации HP: Thermal Logic, Sea of Sensors, и Dynamic Power Capping отсутствующие у конкурентов

iLO3 & Insight Control для завтрашних ЦОДов уже сегодня

Долгая жизнь

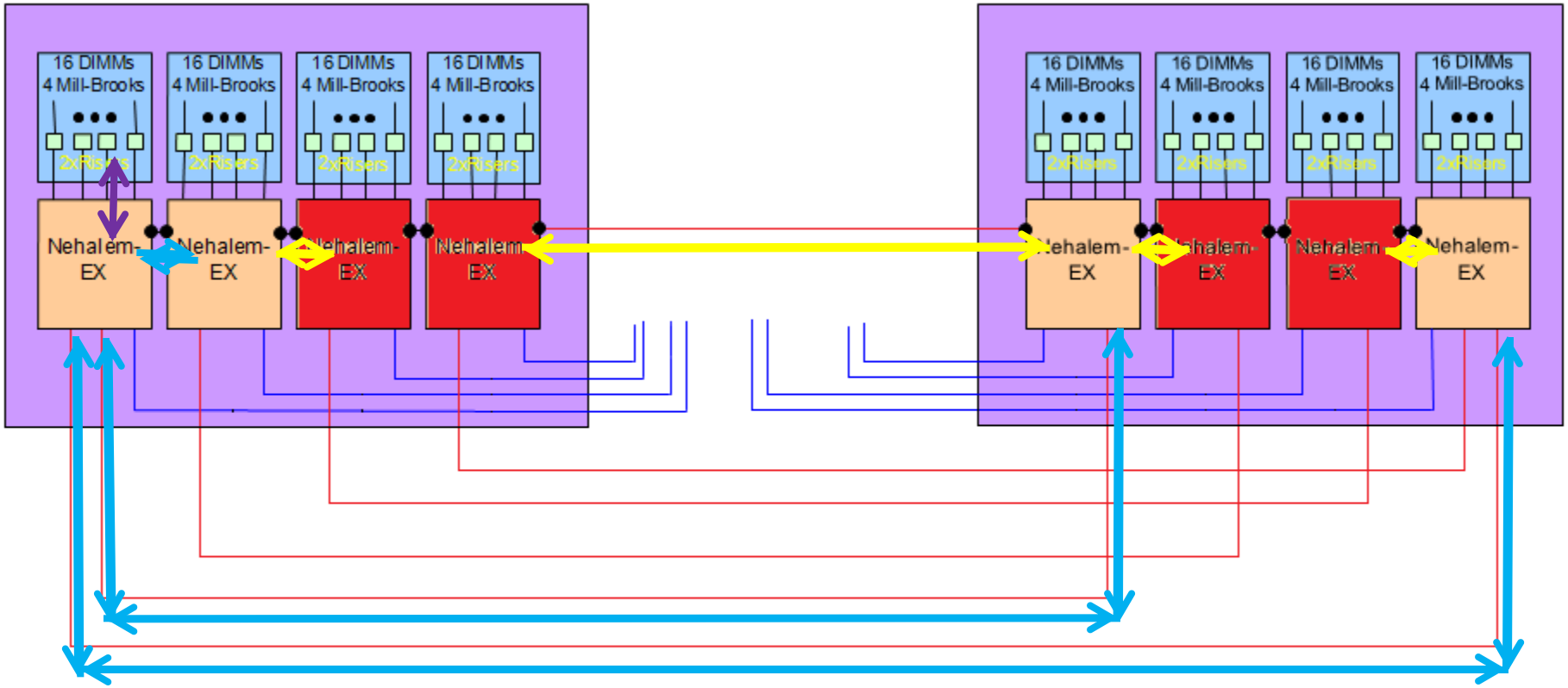
Продажи как минимум до мая 2014 года; 5 лет поддержки

Схема DL980 G7



XNC–контроллер: коммутатор QPI + хранитель информации о кэшах

А если бы XNC не было?



Читаем локальную память (clean)

Ядро ЦП 0 обращается к локальной памяти

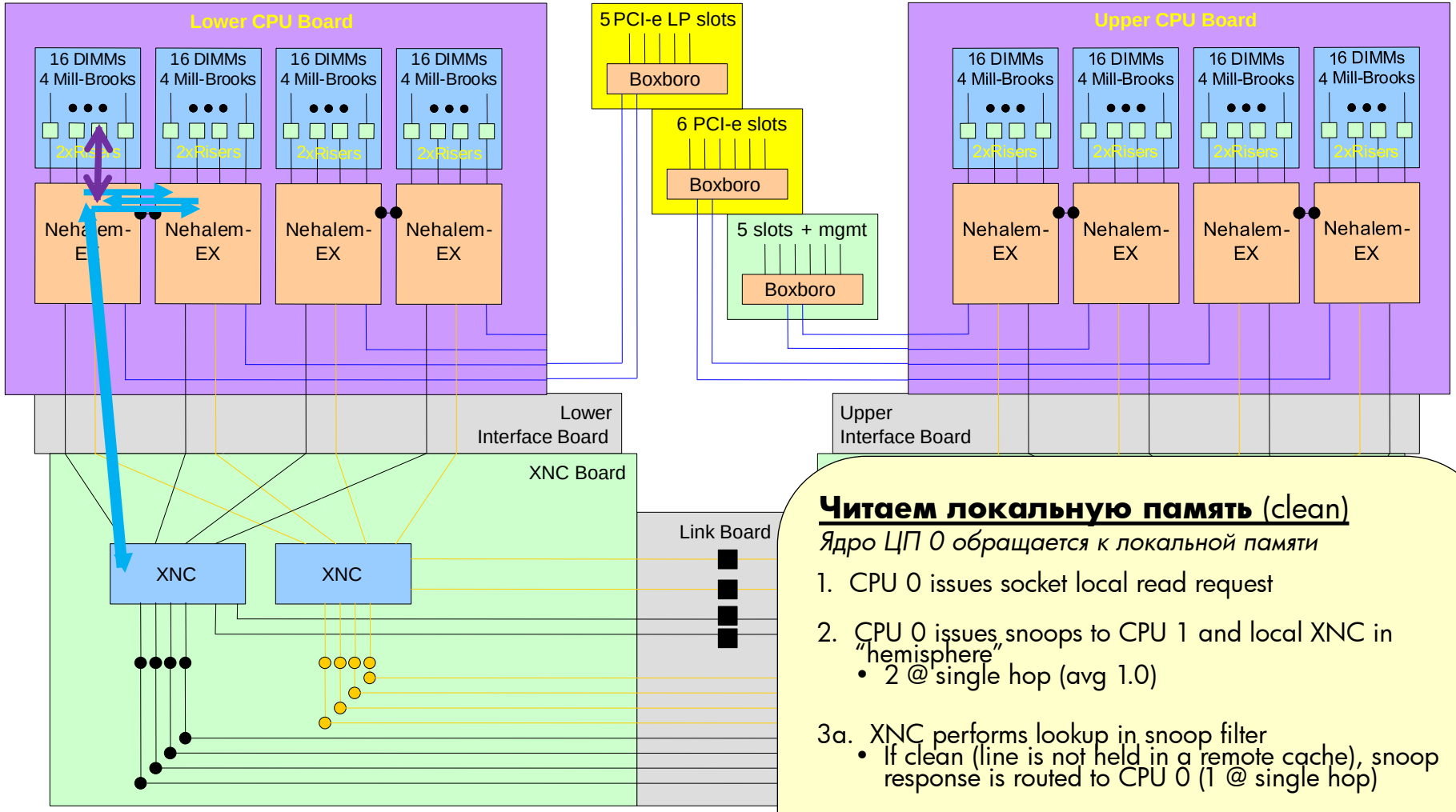
1. CPU 0 issues socket local read request
2. CPU 0 issues snoops to all sockets
   - 4 @ two hops, 3 @ single hop (avg 1.57)
3. Snoop responses routed to CPU 0
   - 4 @ two hops, 3 @ single hop (avg 1.57)
4. Data consumed by core

Наблюдаем SMP overhead:

1.Протокол, обеспечивающий когерентность кэшей плохо масштабируется, свыше 60% пропускной способности связей могут быть заняты запросами когерентности

2.Все ядра испытывают существенную задержку даже при обращении к локальной памяти

Чем помогает XNC?



Читаем локальную память (clean)

Ядро ЦП 0 обращается к локальной памяти

1. CPU 0 issues socket local read request
2. CPU 0 issues snoops to CPU 1 and local XNC in "hemisphere"
   - 2 @ single hop (avg 1.0)

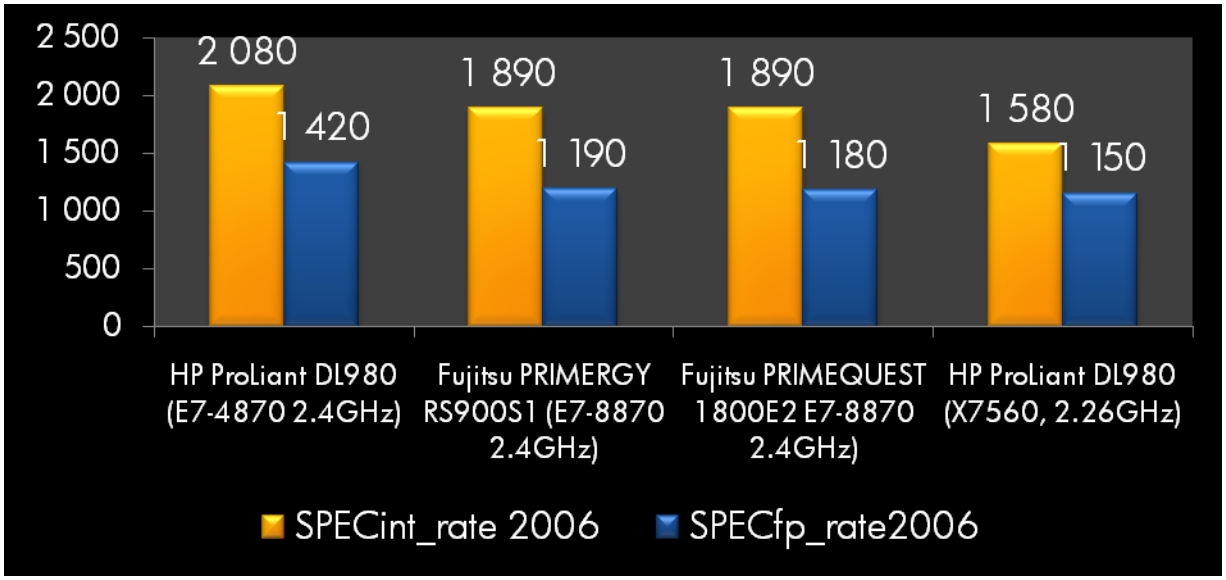
3a. XNC performs lookup in snoop filter
   - If clean (line is not held in a remote cache), snoop response is routed to CPU 0 (1 @ single hop)

3b. CPU 1 performs lookup
   - If clean (line is not held in local cache), snoop response is routed to CPU 0 (1 @ single hop)

4. Data consumed by core

#1 8P x86 single-node SPEC® CPU2006

HP ProLiant DL980: #1 performance in the eight-processor Intel Xeon E7 Family



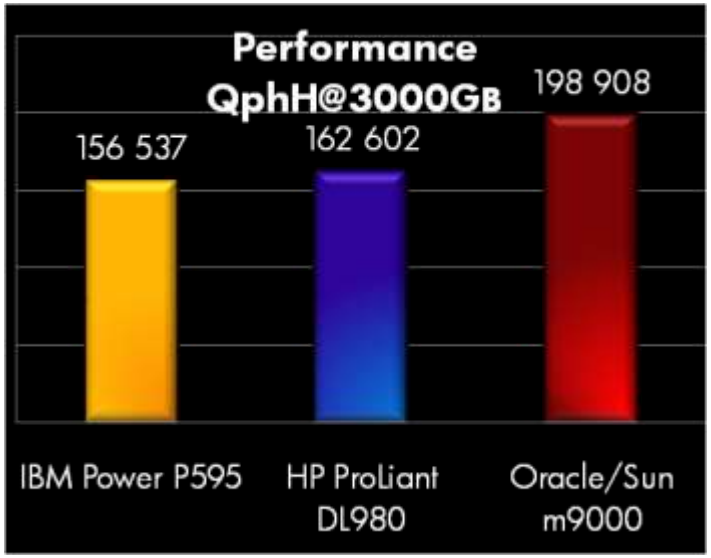
- **10%** better performance in SPECint_rate 2006*
- **20%** better performance in SPECfp _2006*
- **32%** performance gain with the new Intel Xeon E7 Family

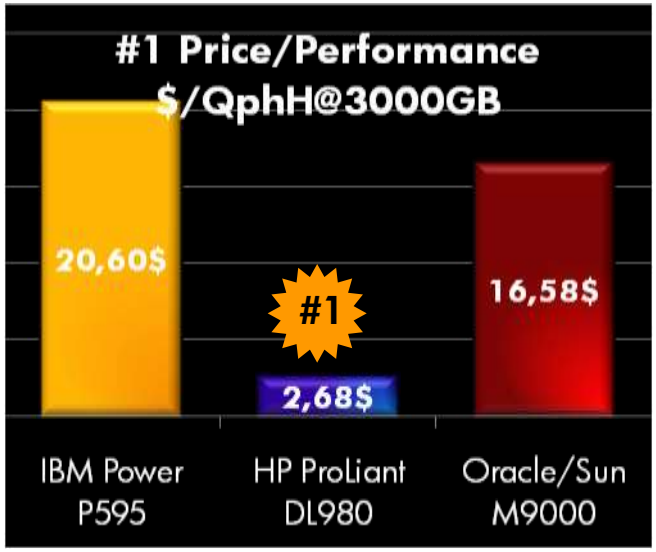
Platform	SPECint_ rate2006	SPECfp_ rate2006	cores/chip	Processor	Memory	OS
HP ProLiant DL980 G7	2,080	1,420	80/8/10	Intel Xeon Processor E7-4870 2.40GHz	1TB	SuSE Linux Enterprise Server 11 (x86_64) SP1
Fujitsu PRIMERGY RX900 S1	1,890	1,190	80/8/10	Intel Xeon Processor E7-8870 2.40GHz	1TB	SuSE Linux Enterprise Server 11 (x86_64) SP1
Fujitsu PRIMEQUEST 1800E2	1,890	1,180	80/8/10	Intel Xeon Processor E7-8870 2.40GHz	1TB	Red Hat Enterprise Linux Server 6.0 x86_64 (SPECint_rate2006); SuSE Linux Enterprise Server 11 (x86_64) SP1 (SPECfp_rate 2006)
HP ProLiant DL980 G7	1,580	1,150	64/8/8	Intel Xeon Processor X7560 2.27GHz	1TB	SuSE Linux Enterprise Server 11 (x86_64) SP1

Note: All results noted were achieved at http://www.spec.org Results as of June 8, 2011

#1 non-clustered TPC-H price/performance @3000GB

HP ProLiant DL980: Extreme Performance at fraction cost





- **1/6th** the price per query of the Oracle/Sun M9000
- **1/8th** the price per query of IBM Power P595
- **1/5th** the physical size of the competition

Platform	Availability	QphH@3TB	USD $/QphH	Processor (chips/cores/threads)	Memory	OS and Database
HP ProLiant DL980 G7	10/13/2010	162,602 QphH @ 3000 GB	$2.68 USD/ QphH@ 3000 GB	Intel Xeon X7560 2.27GHz, (8/64/128)	512GB	Microsoft Windows Server 2008 R2 Datacenter x64, SQL Server 2008 Enterprise Edition x64
IBM POWER 595 Model 9119-FHA	11/04/2009	156,537 QphH @ 3000 GB	$20.60 USD/ QphH@ 3000 GB	Dual-Core IBM POWER 6 – 5.0GHz (32/64/128)	512GB	AIX v 6.1, Sybase IQ Single Application Server Edition v.15.1 ESD #1.2
Sun SPARC Enterprise M9000	04/05/2011	198,907.5 QphH@3000GB	$16.58 USD/ QphH@3000GB	SPARC64 VII 2880 MHz (32/128 cores/256)	512G	Solaris 10, Oracle Database 11g Release 2 Enterprise Edition with Partitioning

Note: All results noted were achieved at www.tpc.org Results as of May 31, 2011

Какой ввод-вывод выдерживает DL980?

ProLiant DL980 G7

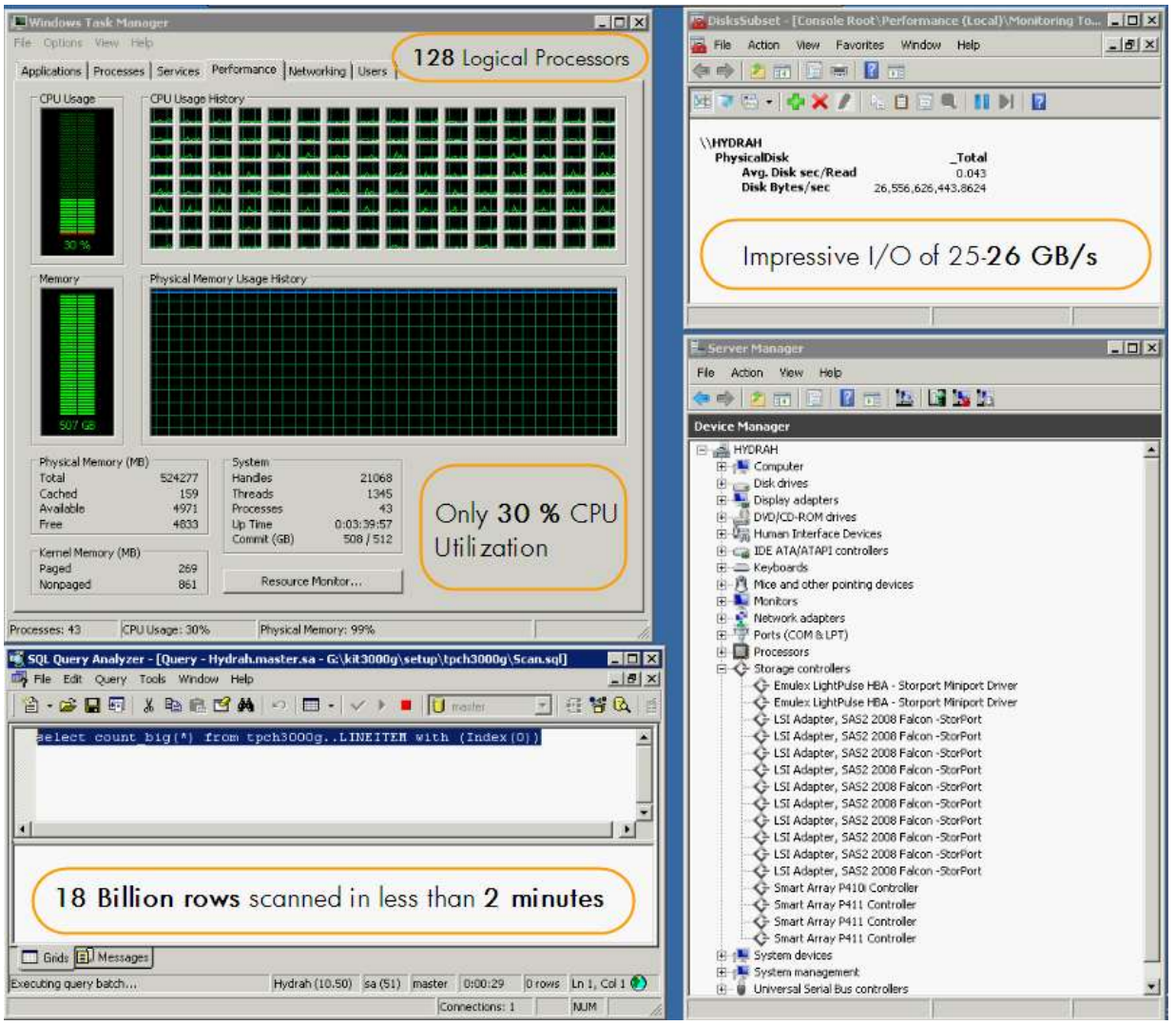
512ГБайт ОЗУ

SQL Index scan:

3 TB Data warehouse
SQL Server 2008 R2

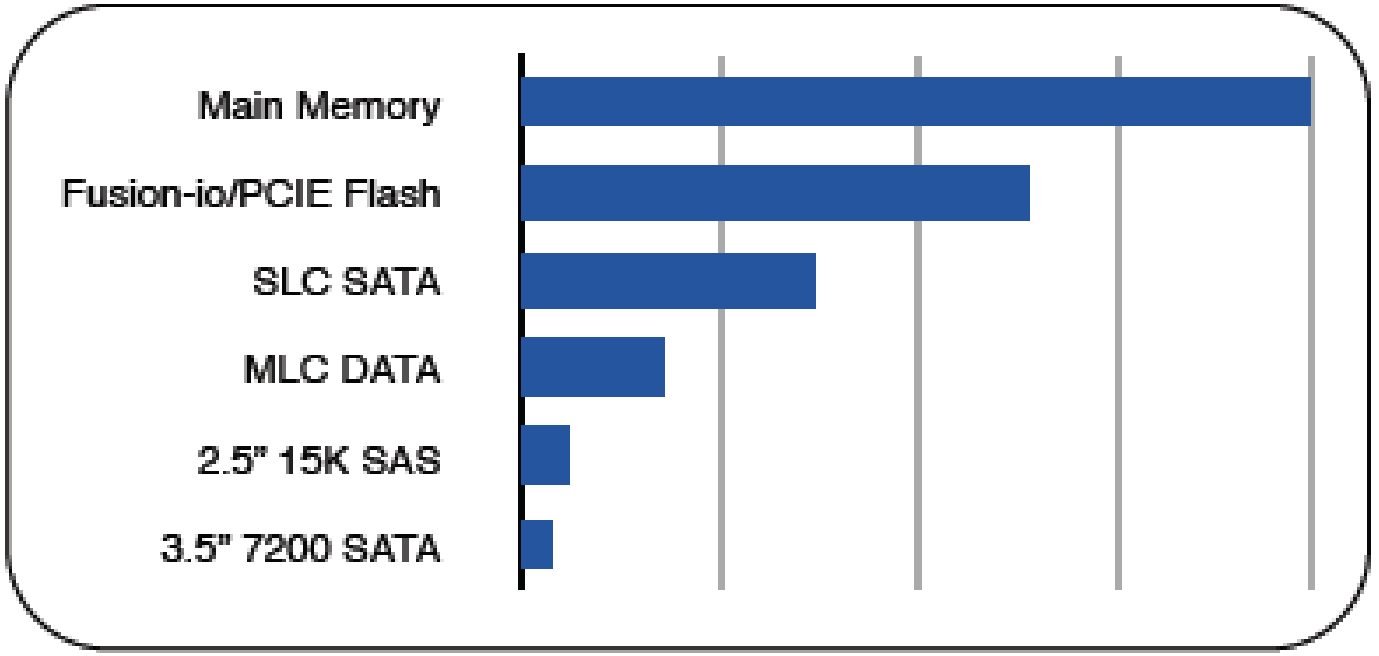
30% загрузки процессоров

25-26 ГБайт /секунду



Ускорим I/O посредством SSD?

Заодно упростим настройку и уменьшим занимаемое место!



* относительная производительность взята с
http://www.vertica.com/2010/08/18/vertica-flash-runaway-performance-at-a-low-price/

Нужно выбрать правильное решение…

PCI IO Accelerator

+ Прямо на PCI – очень большая производительность
+ места в стойке не занимает – внутри сервера
+ удобно для небольших БД и отдельных серверов

- нет аппаратного RAID, поэтому high availability нужно делать программно
- сложности замены или нет горячей замены
- невозможно разделять между серверами

Внешний дисковый массив с SSD

+ Высокая надежность и масштабируемость
+ Разделяемость между серверами через SAN
+ Аппаратный RAID + горячая замена
+ Возможность автоматизированной перекладки наиболее часто используемых данных с дисков на SSD внутри массива

- требует значительных инвестиций
- занимает весьма немало места
- overhead дисковых протоколов

Локальные Solid State Drives

+ Удобный путь использовать SSD
+ быстрее механических дисков
+ Smart Array RAID + горячая замена

- overhead дисковых протоколов
- медленны по сравнению со специализированными флеш-массивами
- не разделяемые между серверами
- живут меньше, чем модули в специализированных флеш-массивах

Специализированные флеш-массивы HP VMA-series

+ Прямо на PCI – лучшая производительность
+ Хорошая масштабируемость для больших (даже >50ТБ) СУБД
+ Аппаратный RAID + горячая замена компонентов
+ Может быть разделяемым через FC SAN gateway
+ Высокая скорость и большой срок жизни одновременно

- офиц. поддержка с прямым PCI только в DL980
- каждые 5ТБ - 10ТБ занимают 3U

Вынеси мозг базам данных

Как обойтись без дисков:
массивы HP VMA

Проблемы, которые нужно преодолеть при содании флеш-массива

(чем флеш-память отличается от дисков)

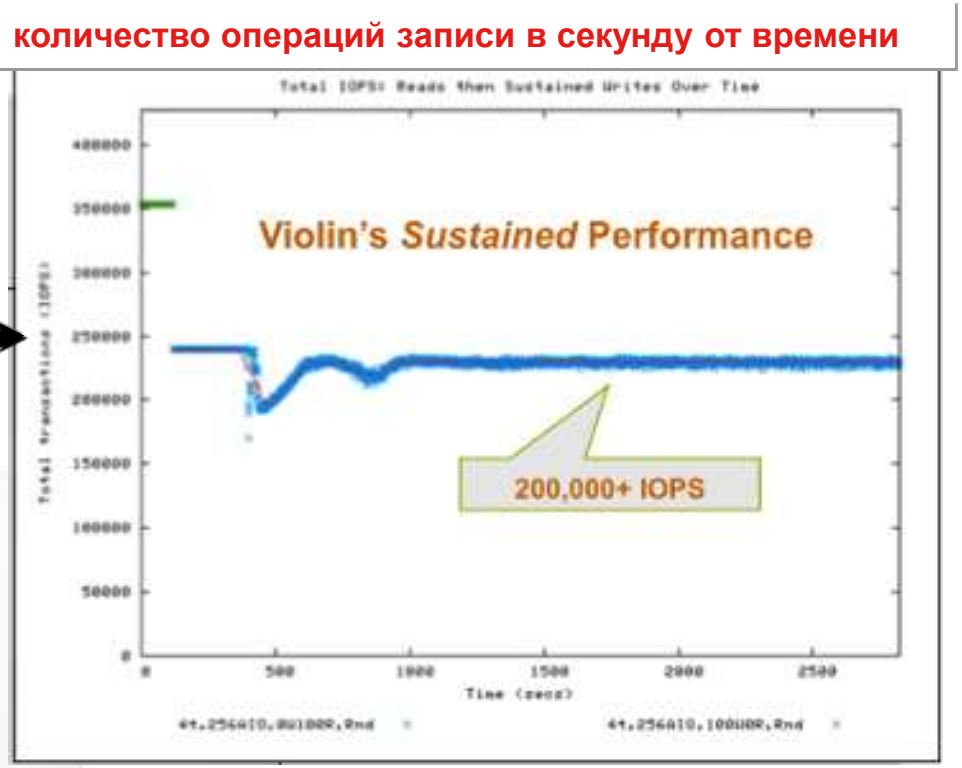
1. Запись на флеш медленне, чем чтение
2. Запись должна быть последовательной в пределах блока 128-256кБ
3. Типичные блоки флеш-памяти больше, чем типичные блоки данных
4. Флеш-блок должен быть стерт, чтобы произвести новую запись
5. Стирание требует значительного времени (миллисекунды) и может блокировать чтение и/или запись на тот же чип(!)
6. Флеш-блок может быть стерт только конечное число раз (физ. износ)
7. Ошибки чтения возрастают по мере увеличения количества чтений
8. Флеш может терять данные (стекает заряд), даже если нет обращений
9. Выйти из строя может не только блок или страница, но и чип целиком

Последние 20 лет алгоритмы RAID развивались именно для механических дисков

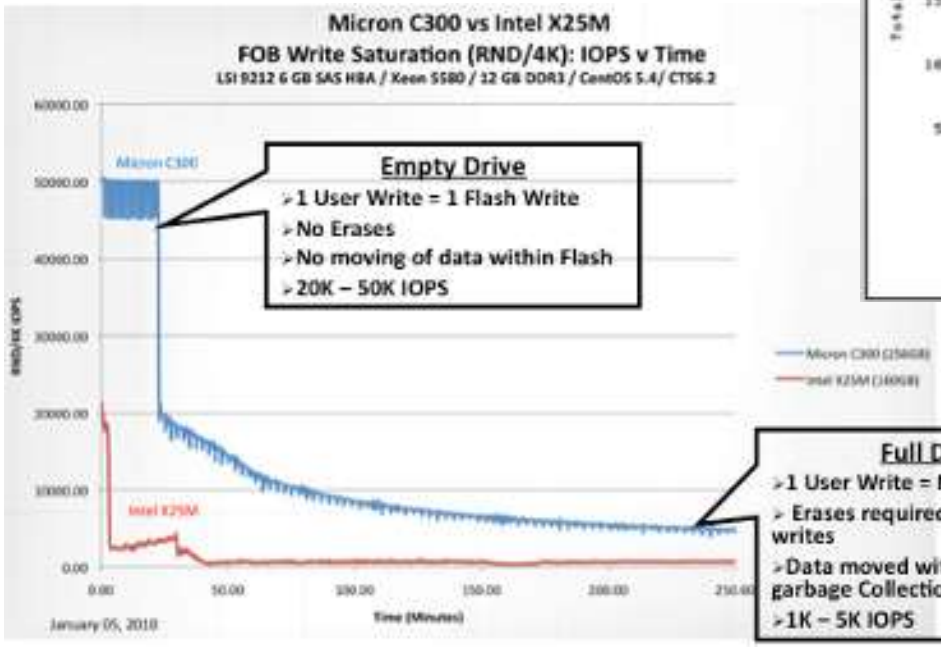
Значит, нужны специальные алгоритмы для флеш-массива

специализированный
флеш-массив (Flash RAID)

количество операций записи в секунду от времени



количество операций записи в секунду от времени



«обычные» SSD

Source: http://www.anandtech.com/tradeshows/showdoc.aspx?i=3712

Как это устроено

Конструкция
упрощает эксплуатацию и
уменьшает стоимость владения

Непревзойденная масштабируемость и плотность для массива энергонезависимой памяти

HP VMA-series Memory Array

- Flash VIMMs — агрегирование микросхем
  - 5ТБ и 10ТБ «сырых» в 3U (VMA3205, VMA3210) - SLC
  - **постоянная скорость записи** (не деградирует со временем)
  - **возможность горячей замены модулей**
- Flash vRAID — защита данных
  - не зависит от механики, в том числе по задержкам
  - 80% эффективности (против 50% у RAID-1)
  - Fail-in-place (самовосстановление, замена крупных модулей)
  - надежность 99.999%
  - срок службы **5 лет** при макс. скорости записи (8ТБ в час)
- Производительность
  - 250 000 IOPS на аппарат (до 2 700 000 IOPS на шкаф)
  - Решение справляется с 200 000 000 транзакций в час
  - Задержка <100 микросекунд (а у дисков в 30 раз больше)
  - Предсказуемая задержка (а у дисков разброс из-за механики)
  - 1,4ГБайт/сек на аппарат (свыше 12ГБайт/сек на шкаф)
  - Подключается напрямую по PCIe к серверу

Как добиться производительности?

Транзакционная система (OLTP)

Конвергентная Инфраструктура

OLTP Application Servers

Приложения:
- транзакционные
- аналитические

HP ProLiant DL980 G7

HP VMA-series Memory Array

HP High Performance Database Solution

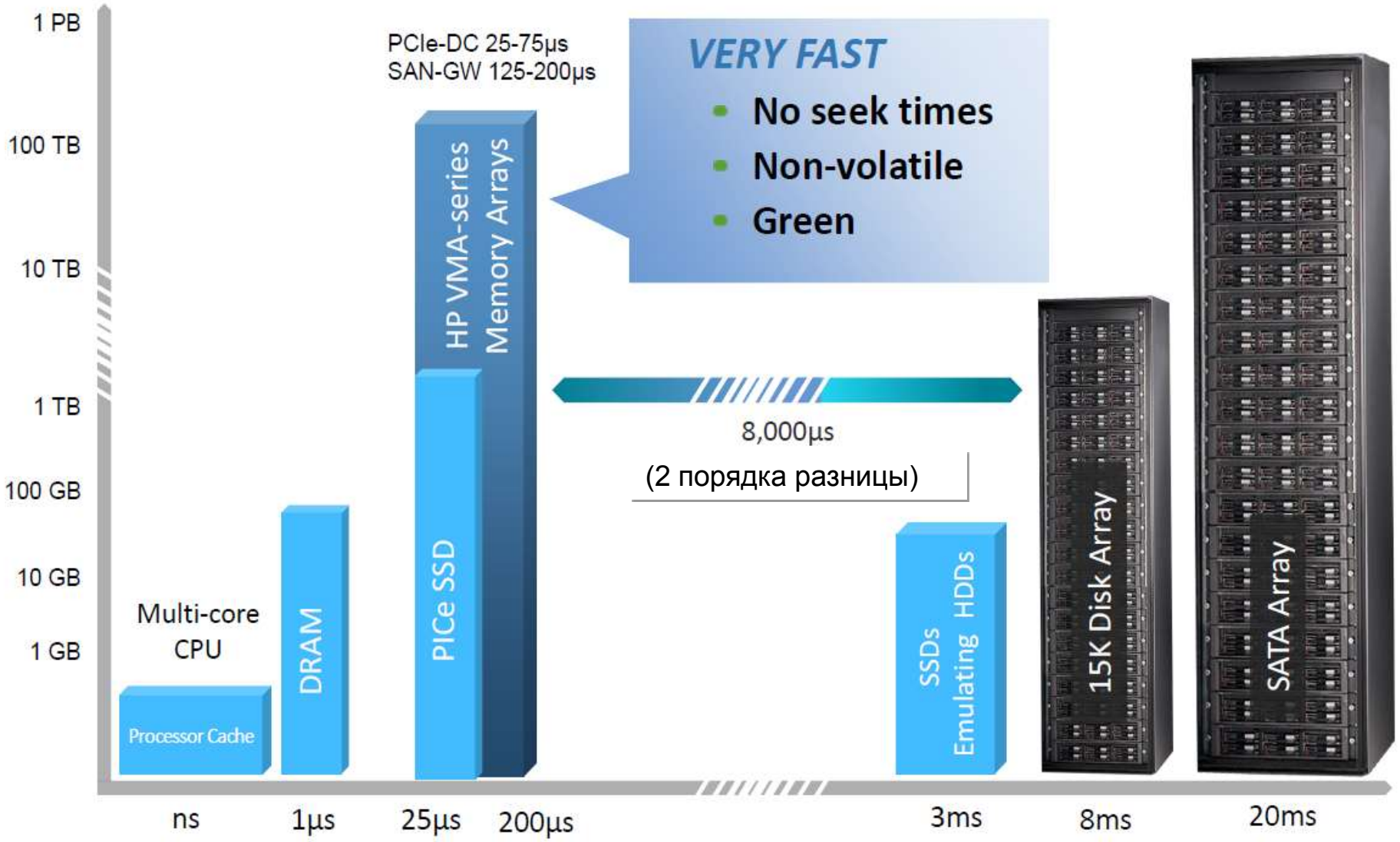
СУБД

- **и OLTP, и аналитика**
- **в 8-10 раз быстрее, чем на дисках!**

HP Storage Solutions

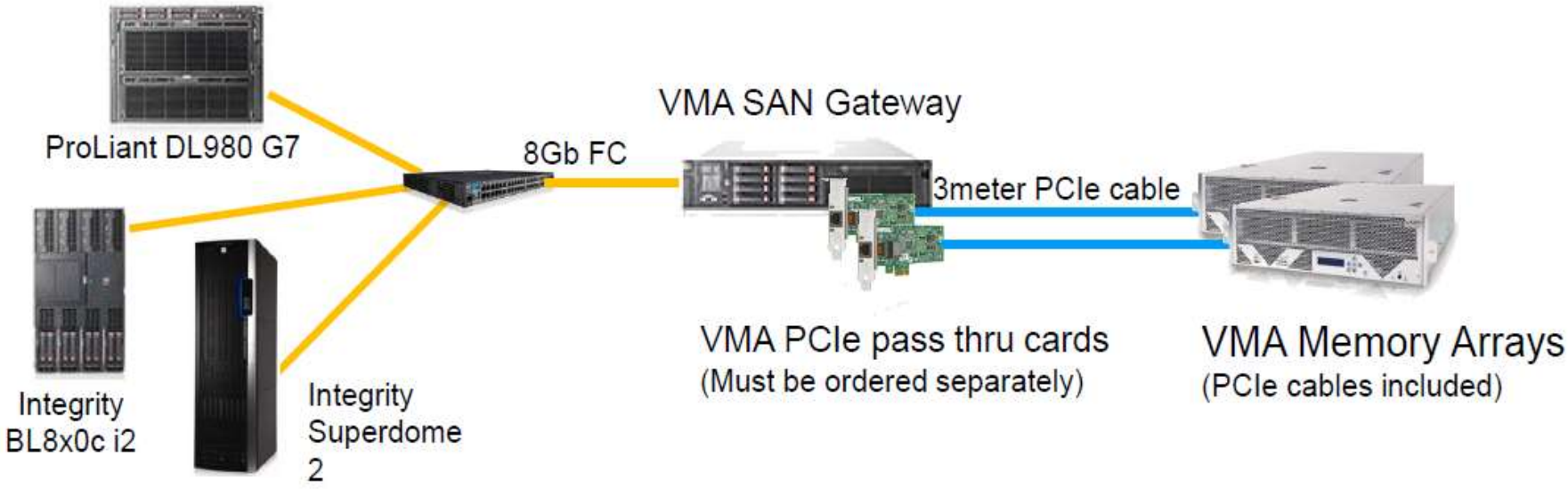
Диски для:
- резервного копирования
- катастрофоустойчивости
- дублирования

Как добиться скорости ввода-вывода?



Время отклика (задержка доступа)

Как подключить VMA в SAN?



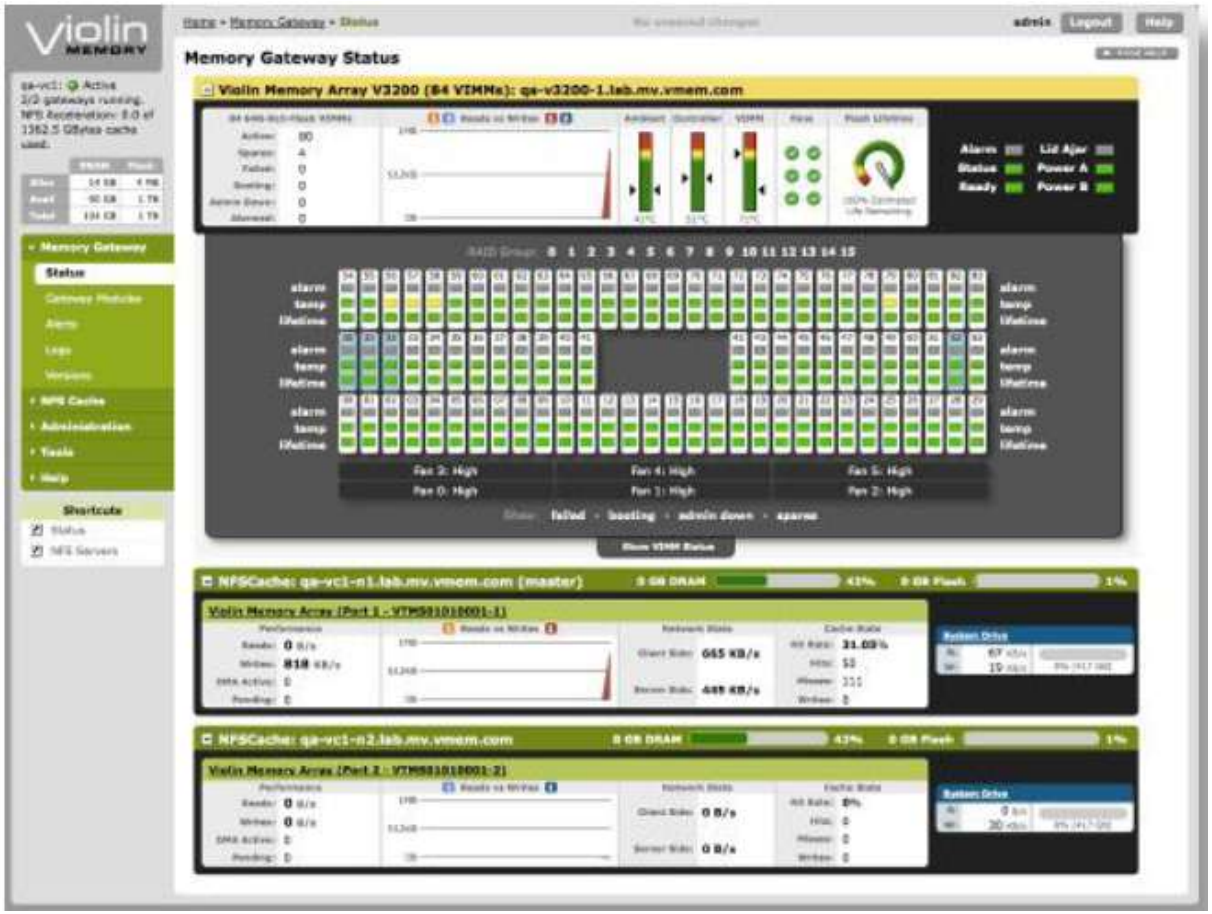
- SAN gateway appliance с предустановленным ПО и картами 8Gbit FC
- До 2-х VMA на каждый gateway
- Windows на DL980
- HP-UX на Integrity
- Возможность организации разделяемого пула ресурсов

Как управлять массивами VMA в сети SAN?

интуитивно понятный интерфейс в SAN Gateway отображает:

- статус системы
- износ флеш-модулей
- конфигурацию
- безопасность
- производительность

возможен также доступ через командную строку



DL980 + VMA = Эффективность СУБД

Преимущества:

- Эффективные транзакционные системы + аналитика для очень больших нагрузок любого типа
- Увеличение емкости хранения сервера СУБД без потери производительности и без усложнения СХД.
- Уменьшение стоимости владения поскольку меньше энергии потребляет и меньше места занимает, чем дисковый массив.
- Уменьшение времени внедрения с помошью референсных конфигураций, подготовленных HP.
- До 80ТБайт флеш-памяти на сервер
- До 4ТБ оперативной памяти на сервер
- Виртуально неограниченный рост: серверы можно объединять в кластеры/пулы средствами виртуализации, средствами ОС и средствами СУБД

Тех. поддержка

Оптимальная: оптимальная производительность

- HP ProLiant Server Hardware Installation
- HP Installation and Startup Service for Insight Control Software
- 3-Year HP Critical Advantage

Стандартная: высокий уровень надежности

- HP ProLiant Server Hardware Installation
- HP Installation and Startup Service for Insight Control Software
- 3-Year HP 6 hour Hardware Support Onsite Call-to-Repair Service
- 3-Year HP 24x7 Software Support for Insight Control

Базовая: минимально рекомендованный уровень

- HP ProLiant Server Hardware Installation
- HP Installation and Startup Service for Insight Control Software
- 3-Year HP 24x7 4 hour Response, Hardware Support Onsite Service
- 3-Year HP 24x7 Software Support for Insight Control

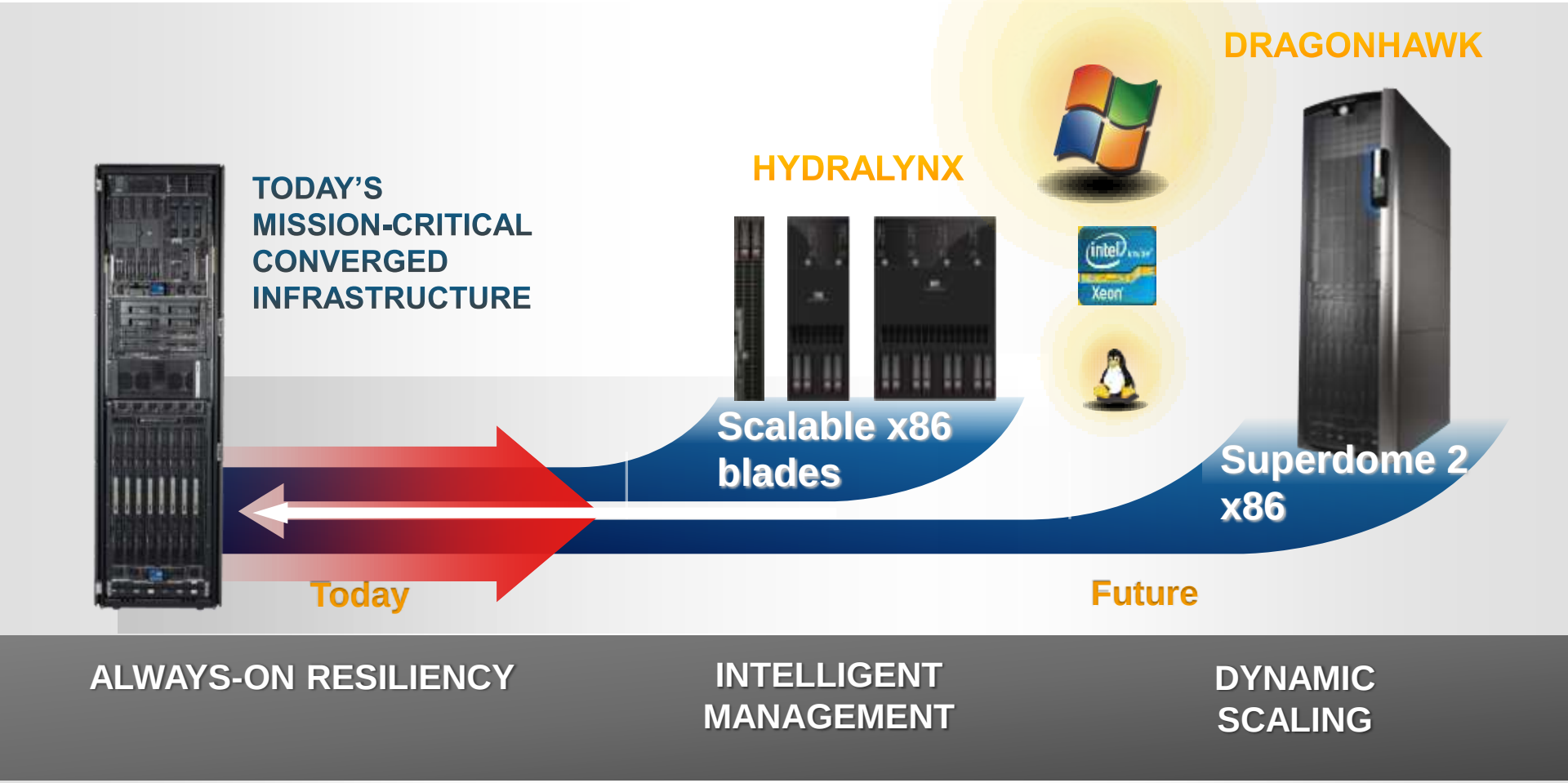
Интегрированные решения на базе референсных конфигураций:

- уменьшают риски
- повышают эффективность

Позиционирование HP Integrity и ProLiant

- HP Integrity – там где простой системы стоит дорого
  - Максимальная живучесть комплекса: кластерные и катастрофоустойчивые системы
  - Максимальная надежность одной машины: SD2
  - Максимальная гибкость и масштабируемость: виртуализация+кластеризация
  - Инфраструктура для долговременной эксплуатации
- HP ProLiant – там где требуется экономия и стандартизация на x86
  - Надежная инфраструктура, оптимизированная по цене/производительности
  - Стандартные опции и приложения x86
  - Широчайший стек решений для платформы x64
  - Универсальный инструментарий для внедрения, управления и сопровождения программно-аппаратных комплексов

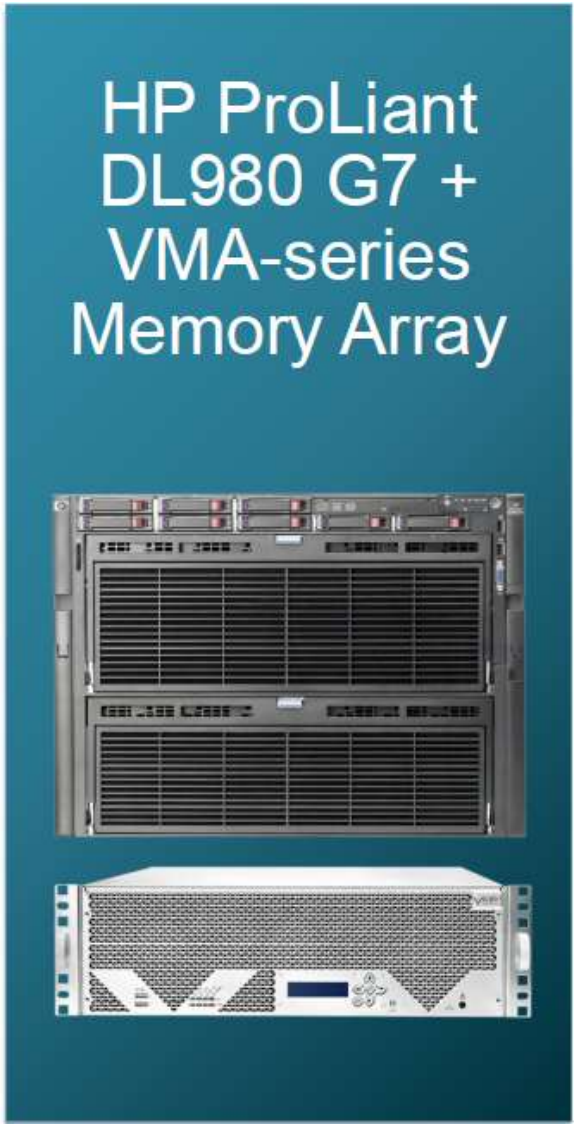
Развитие систем x86 для Больших Данных



Проект «Одиссей»: разработка такой платформы в течение 2-х лет

Итоги

HP Proliant DL980 G7 + VMA - это:



- единое решение, продаваемое и поддерживаемое HP = экспертиза от вендора с поддержкой уровня mission critical services,

- возможность решения широкого круга задач,
- упрощение обслуживания и настройки вычислительного комплекса,
- повышение производительности СУБД на порядок за счет уменьшения времени отклика ввода-вывода

- уникальная архитектура, обеспечивающая:
  - сбалансированную масштабируемость
  - надежность и самовосстановление
  - эффективность

СПАСИБО! / THANK YOU