



Опыт работы и оптимизации ВЦ

Алексей Нечаев
agn@largeo.com
системный инженер

Введение

Наземная съемка 3D	Морская съемка 3D
Время записи, $T = 6$ sec	Время записи, $T = 9$ sec
Дискретизация, $dt = 2$ ms	Дискретизация, $dt = 2$ ms
Количество сэмплов, $N = 3000$	Количество сэмплов, $N = 4500$
Объем сэмпла, $v = 16$ bit	Объем сэмпла, $v = 16$ bit
Объем трассы, $V = 6$ кВ	Объем трассы, $V = 9$ кВ
Средний размер съемки 300 – 500 Gb	Средний размер съемки 7 - 10 Tb
Количество трасс 30 – 80 млн.	Количество трасс 0.8 – 1 млрд.

Данные для расчетов

Плюсы: все расчеты слабосвязанные

Минусы: огромное количество мелких блоков (трасс)

Опыт использования больших машин

Sun Fire 6800, SGI Altix 350, 330, серверы
на процессорах Intel Itanium2

Плюсы: дисковые системы на быстрых
интерфейсах, общее поле памяти

Минусы: дорогая масштабируемость

Вычислительные кластеры

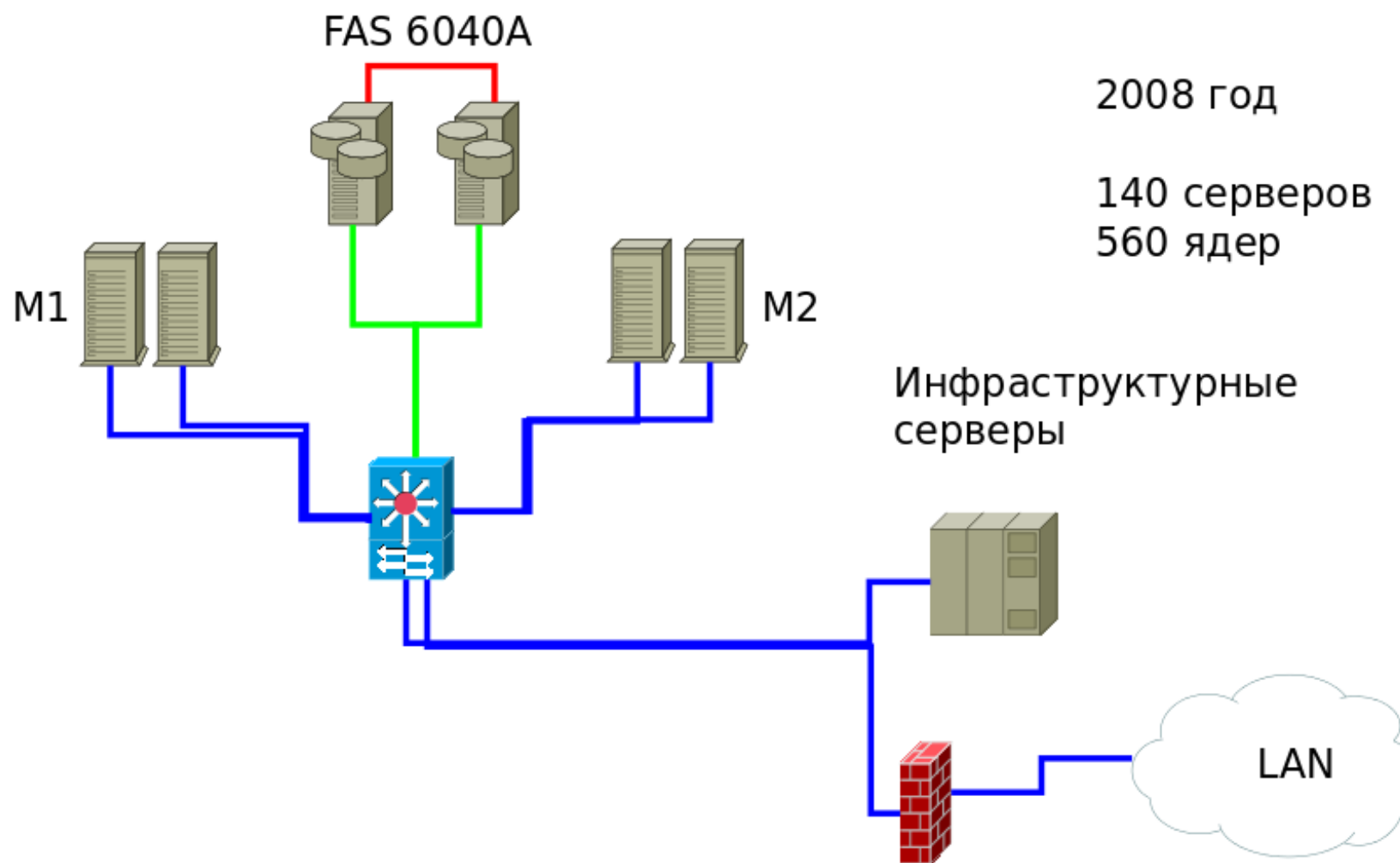
- Смена “образа мышления”: множество кластеров.
- Распределение кластеров между проектами
- Конкуренция за дисковое пространство и дисковый ввод-вывод
- Основная цель: любой проект доступен на любом кластере

Ближайшее будущее:

Морские проекты 2-3 тыс. кв. км

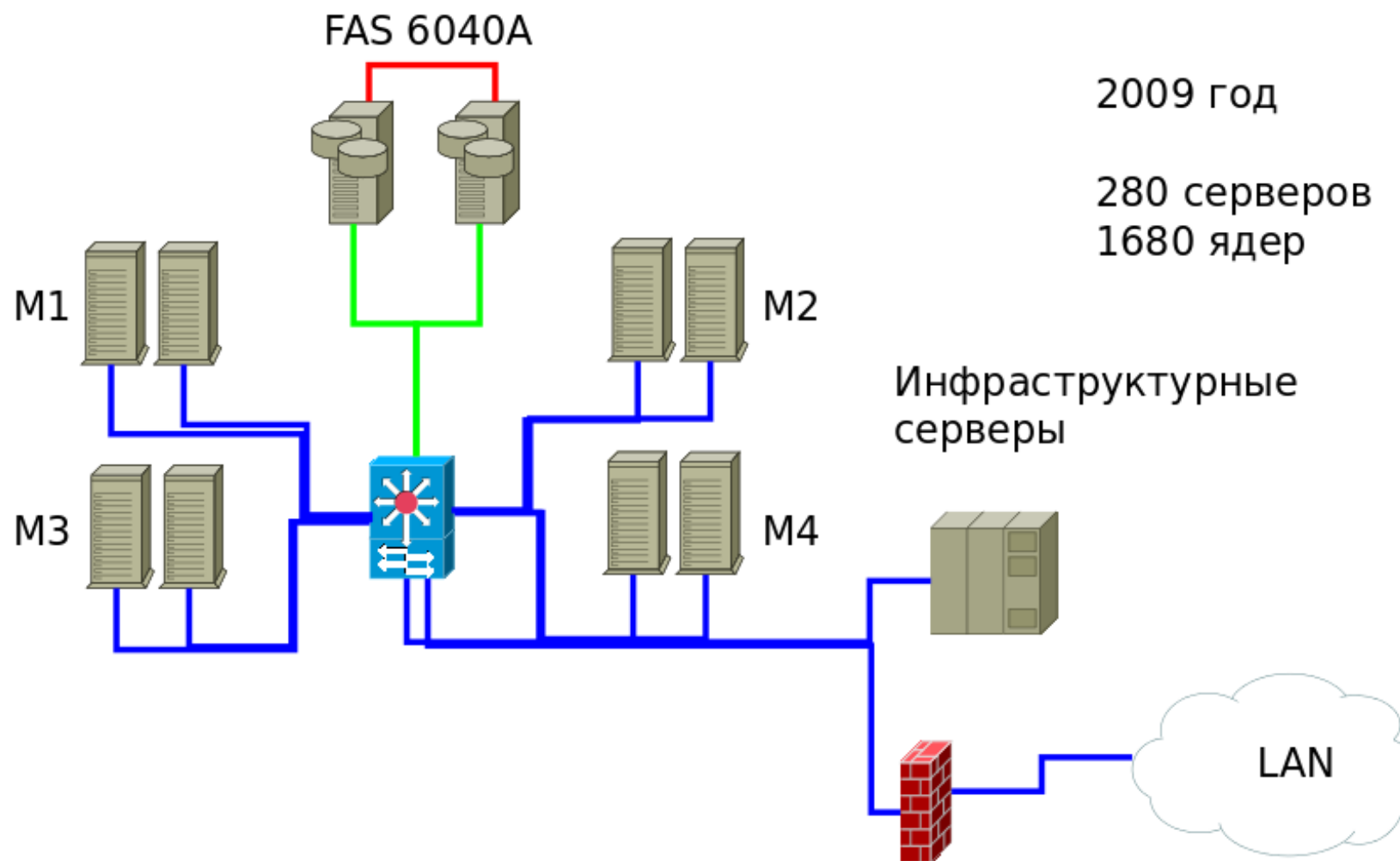
- Объем данных: 20Тb
Количество трасс: 4,5 млрд.
- 3000 - 4500 процессорных ядер
скоростные системы хранения данных

Масштабируемый ВЦ



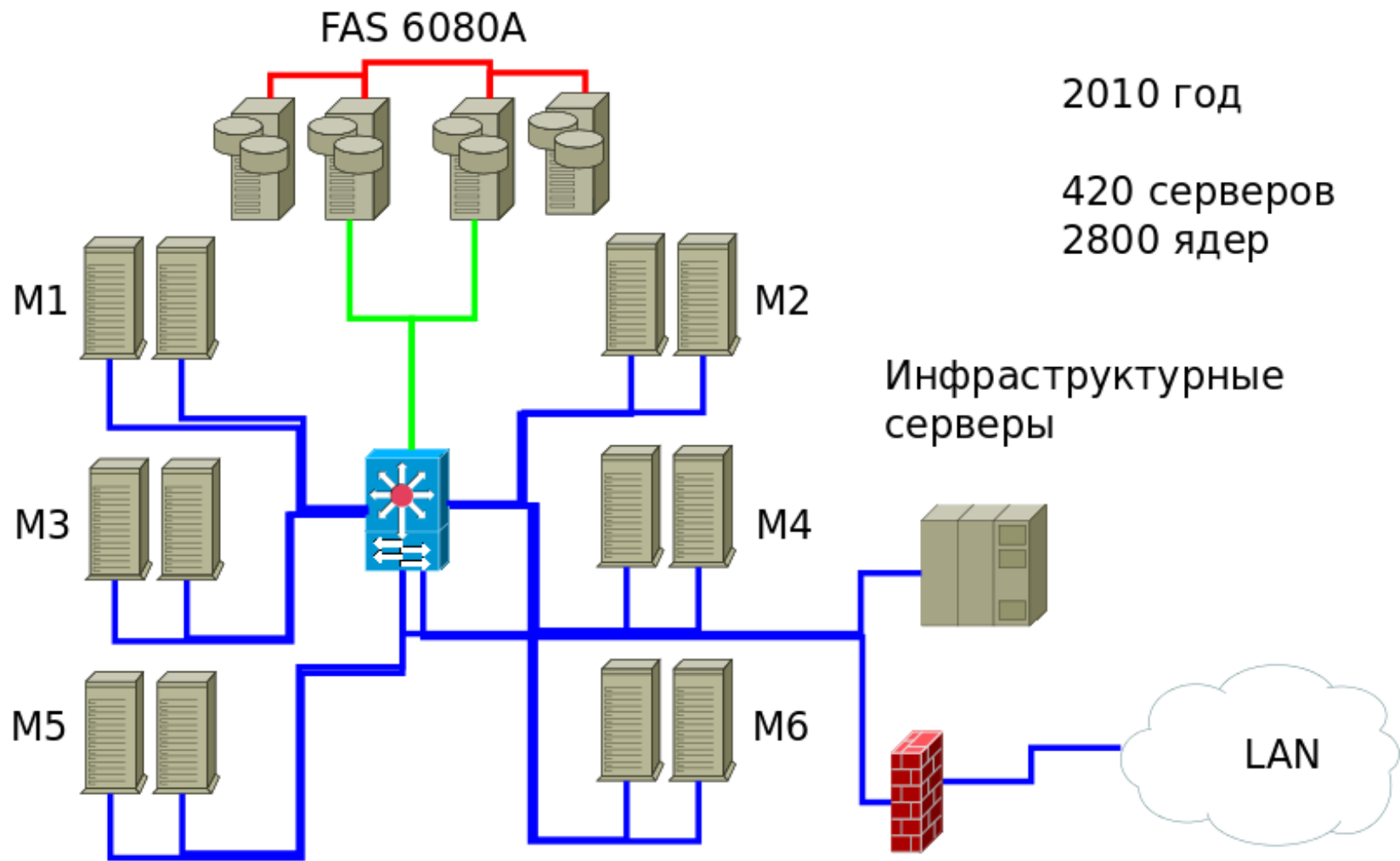
Mx - кластер из 70 узлов

Масштабируемый ВЦ



Мх - кластер из 70 узлов

Масштабируемый ВЦ



2010 год

420 серверов
2800 ядер

Mx - кластер из 70 узлов

Организация проектов. Взгляд со стороны геофизиков

	Project 1	Project 2	Project 3
M1	X		
M2		X	
M3			X
M4			X
M5			X
M6			X

Организация проектов. Взгляд со стороны техпроцессов

	Processing	PreSTM/SDM	SRME	DWDM
M1	X			
M2	X			
M3		X		
M4		X		
M5			X	
M6				X

	PRJ1
	PRJ2
	PRJ3

Организация проектов. Взгляд со стороны СХД

	Zhukov	Bagration
net_project3_vol1	aggr1	
net_project3_vol2		aggr1
net_project3_vol3	aggr2	
net_project3_vol4		aggr2
net_project3_vol5	aggr3	
net_project3_vol6		aggr3
net_project3_vol7	aggr4	
net_project3_vol8		aggr4
...		
net_project3_volN	aggrN	
net_project3_volN+1		aggrN

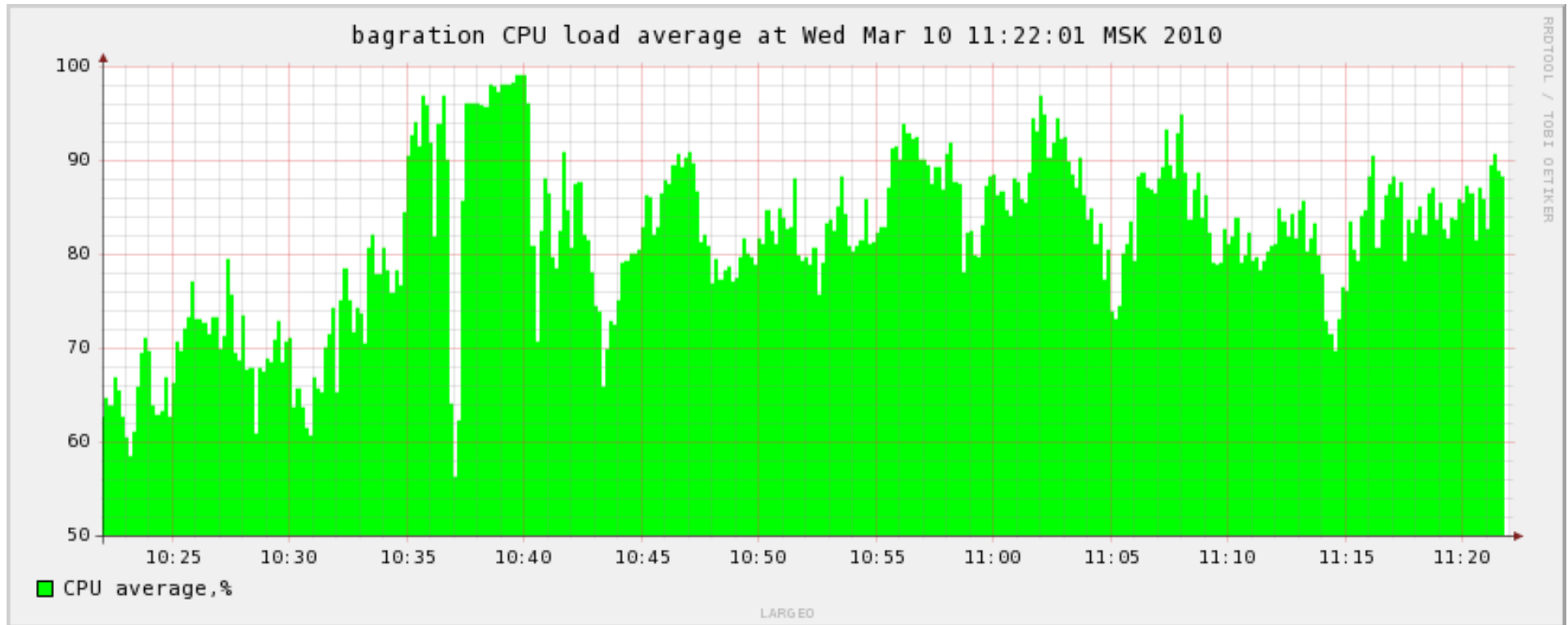
Распределение данных

- **Дисковые полки**
16 дисков FC 15K RPM
- **“Агрегаты”**
48 дисков FC 15K RPM
- **WAFL**
Write Anywhere File Layout
- **Тома данных**
Пространство имен NFS, объемом 2-3 Tb

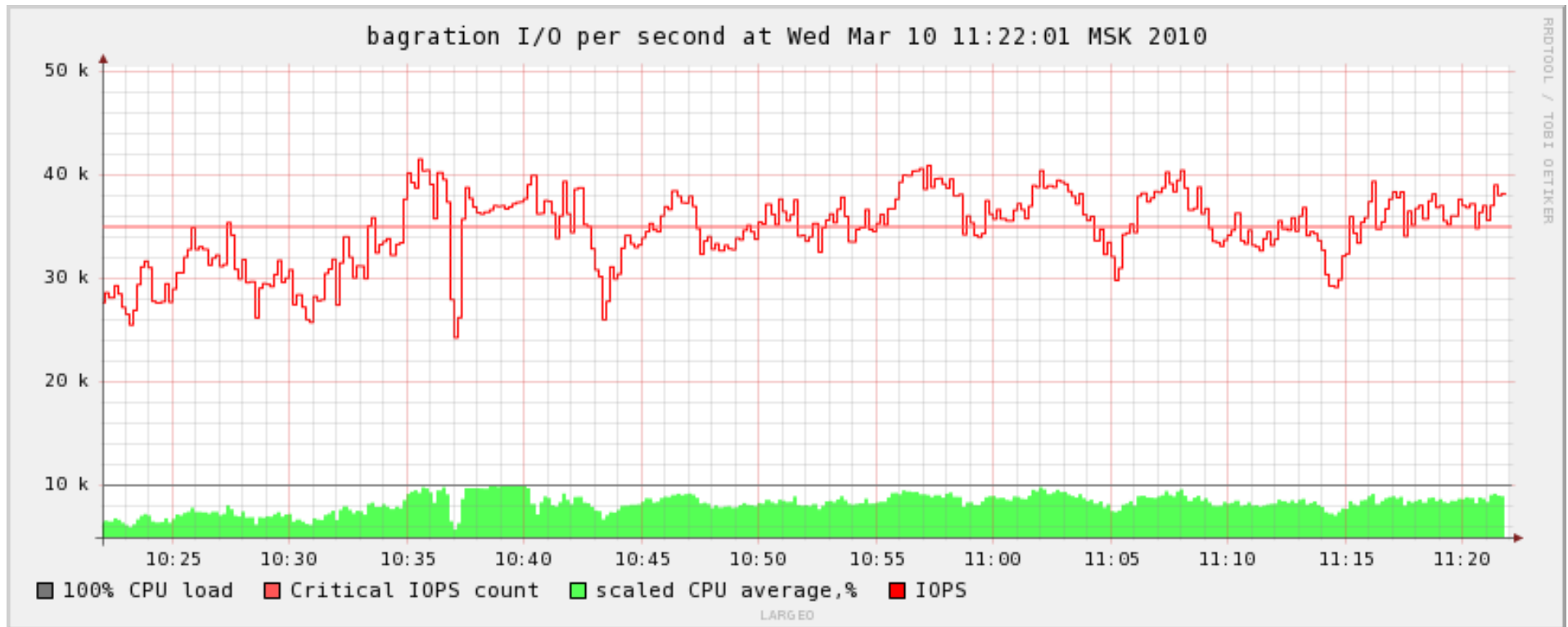
Управление и мониторинг

- Управление серверами: IBM xCAT
- Управление пользователями: NIS
- Управление заданиями: CJM, Torque [PBS]
- Мониторинг серверов: Ganglia, SNMP, IPMI

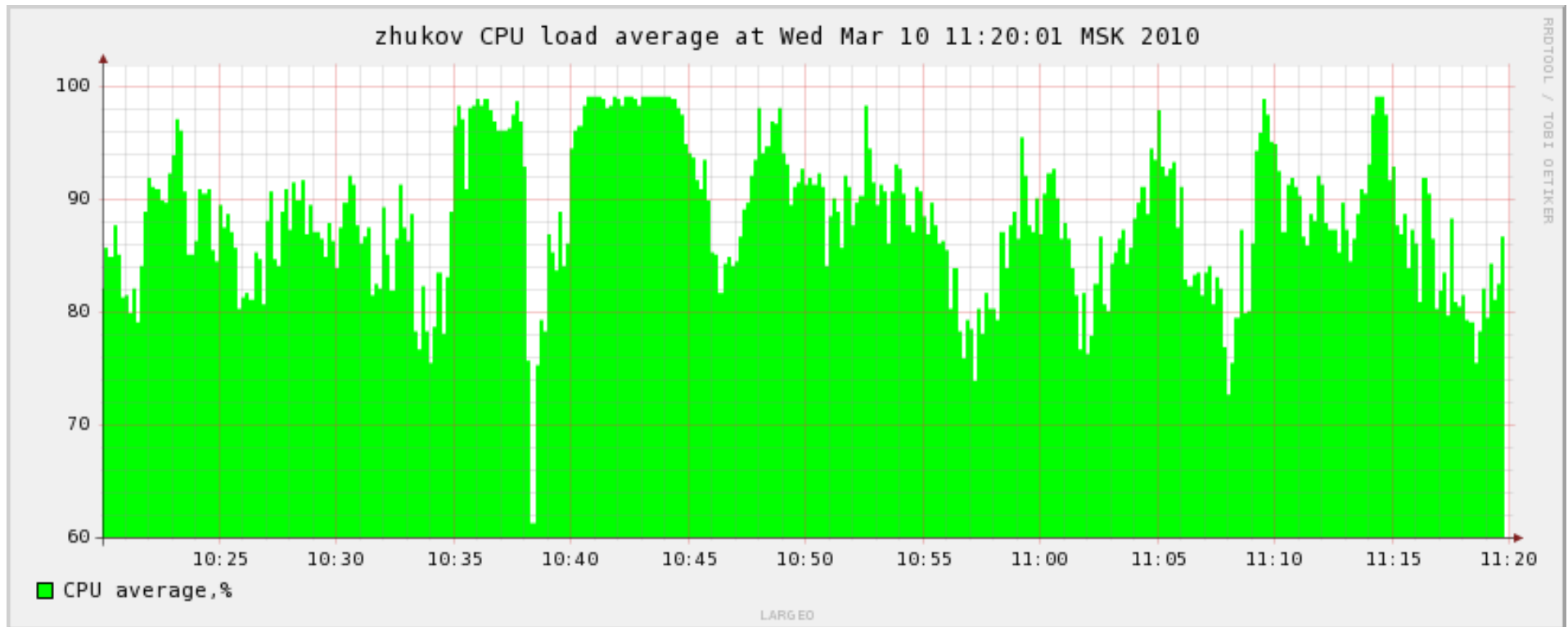
Хорошо нагруженная СХД



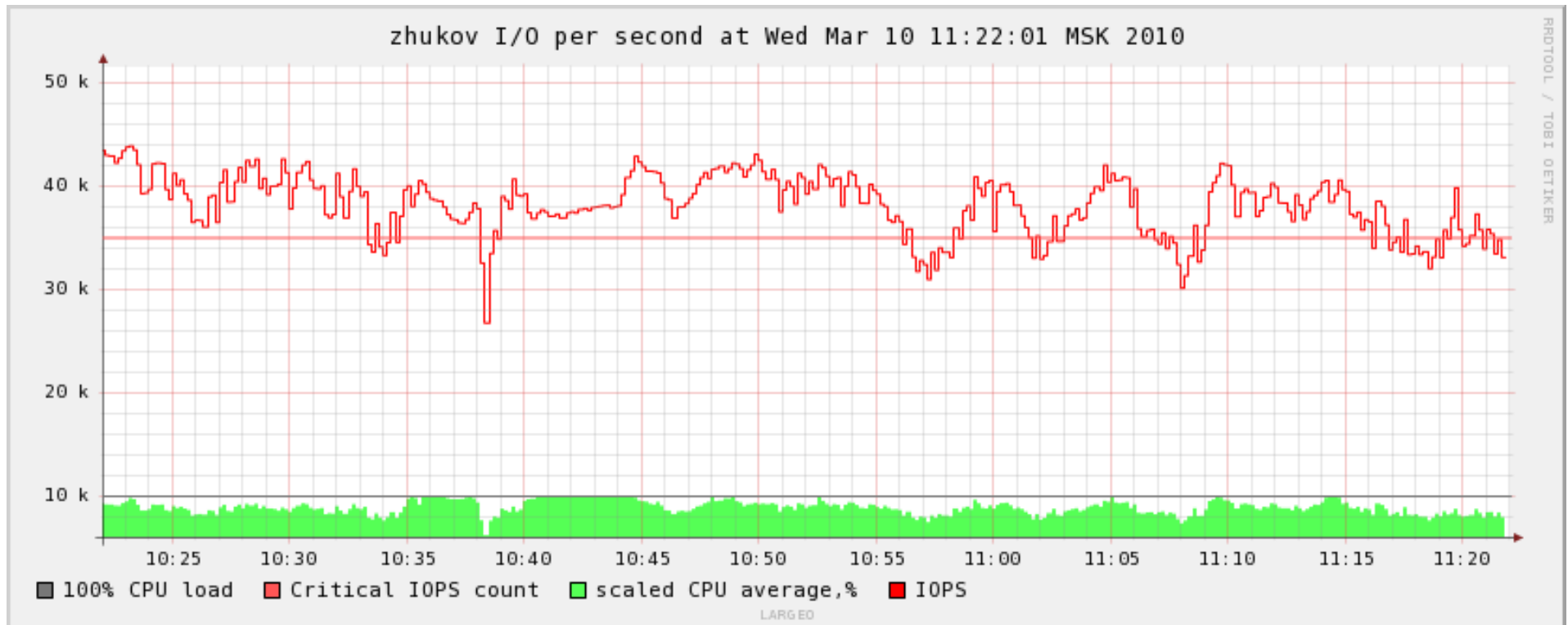
Хорошо нагруженная СХД



Хорошо нагруженная СХД



Хорошо нагруженная СХД



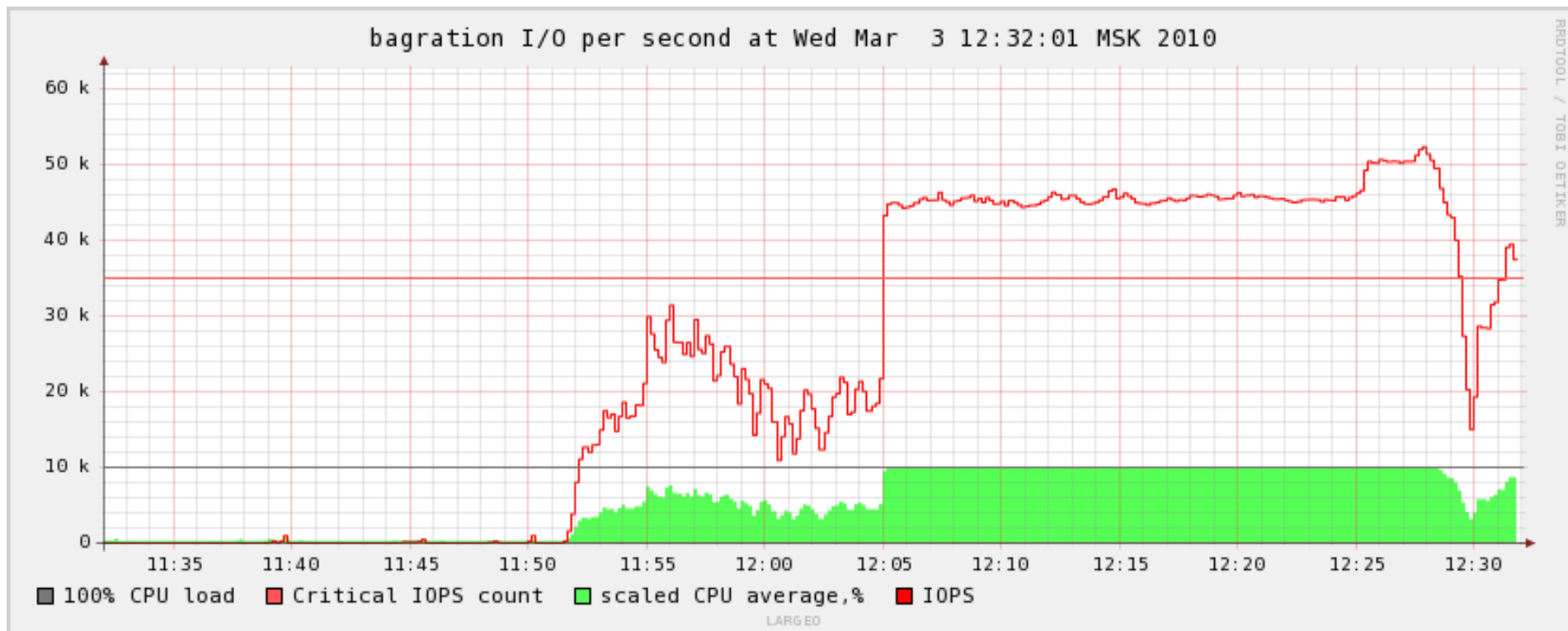
Инструменты для анализа производительности СХД

- `sysstat -us 1`
- `sysstat -m 1`
- `stats show vfiler nfsv3`
- `stats show vfiler volume`

`options nfs.per_client_stats.enable on`

- `nfsstat -l 10`
- `nfsstat -h`

Предельно нагруженная СХД DWDM, 280 CPU cores



Итог:

“Всё уже придумано до нас.”

- Ядро сети: коммутатор L3 с производительной матрицей
- Системы хранения данных с быстрым NFS-ом
- “Типовые” серверы
- Открытые софтверные компоненты управления и мониторинга
- Максимум внимания прикладному ПО