

Российская академия наук
Уральское отделение
Институт механики сплошных сред УрО РАН



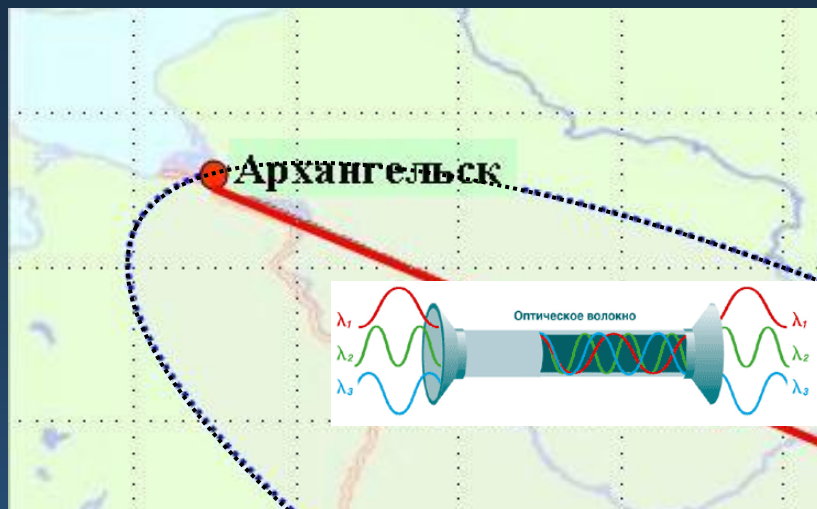
Распределенный интерконнект: структура параллельных каналов и инструментарий исследования транспортных потоков в них

Масич А.Г.	mag@icmm.ru	+7 (342) 2932739
Масич Г.Ф.	masich@icmm.ru	+7 (342) 2378376
Сидоров И.А.	sidorov@icmm.ru	+7 (342) 2378395

Международная суперкомпьютерная конференция
Научный сервис в сети Интернет. Все грани параллелизма
23-28 сентября 2013

- Сегодня пропускная способность сетей - не дефицитный ресурс
- Повсеместное развитие национальных и региональных R&E Lambda сетей привел к созданию глобальной Lambda системы
 - ✓ Сообщество GLIF (Global Lambda Integrated Facility), 3rd LambdaGrid Workshop Рейкьявик (Исландия), август 2003
- GLIF разделяет и продвигает **Lambda Grid** парадигму распределенных вычислений
 - ✓ центральный архитектурный элемент это скоростные оптические сети,
- Эта парадигма основывается на использовании параллелизма, как и в суперкомпьютерах десятилетие назад
- В такой модели параллельные лямбда потоки являются системообразующим компонентом
- И проблема состоит в создании аппаратной и программной архитектуры для понимания этой высокоуровневой вычислительной модели и оценки ее преимуществ по сравнению с традиционной моделью кластерных вычислений

Сети: Инициатива GIGA UrB RAS (2008)



Объединяет научные центры и ВУЗы в городах

- Архангельск (Архангельский научный центр)
- Сыктывкар (Коми научный центр)
- Ижевск (Удмуртский научный центр)
- Пермь (Пермский научный центр)
- Екатеринбург (УрО РАН)

Использует технологии

- ✓ Темные волокна $L \sim 2340$ км
- ✓ DWDM: $16\lambda \times 10-100$ Гбит/с
- ✓ Скорость: $0,16-1,6$ Тбит/с

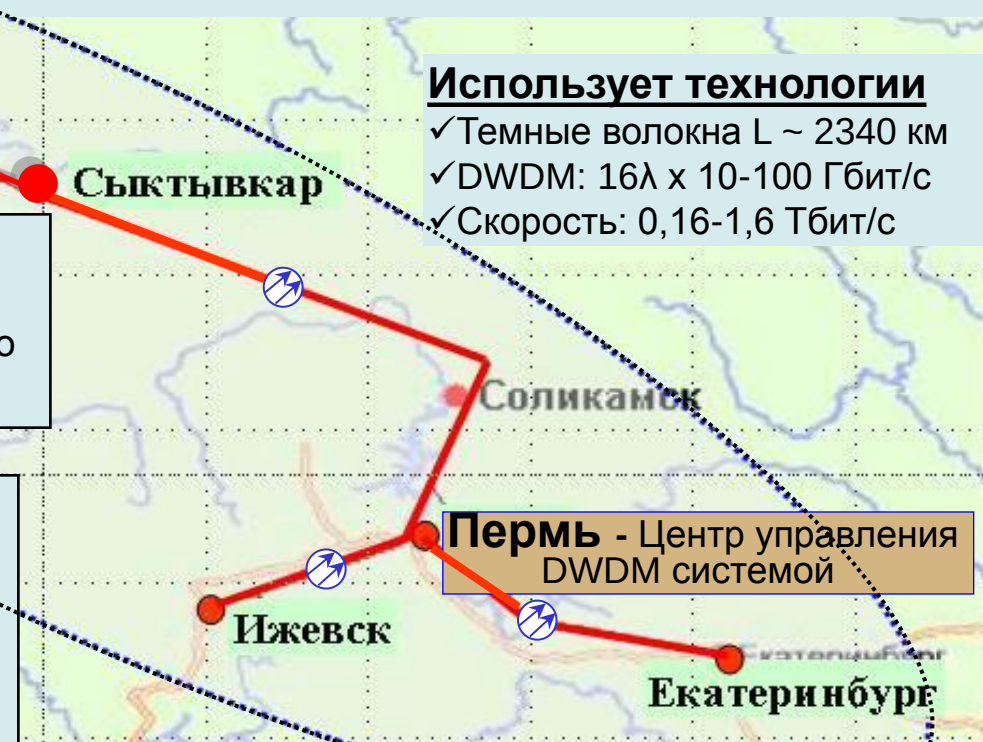
Экономически эффективна:

- предлагаемое решение (40 Гбит/с) – 345 млн. руб. разово
- возможная аренда 0,2 Гбит/с – 400 млн. руб. ежегодно

345 млн.руб разово << 400 млн.руб ежегодно

Инвестиционная защищенность

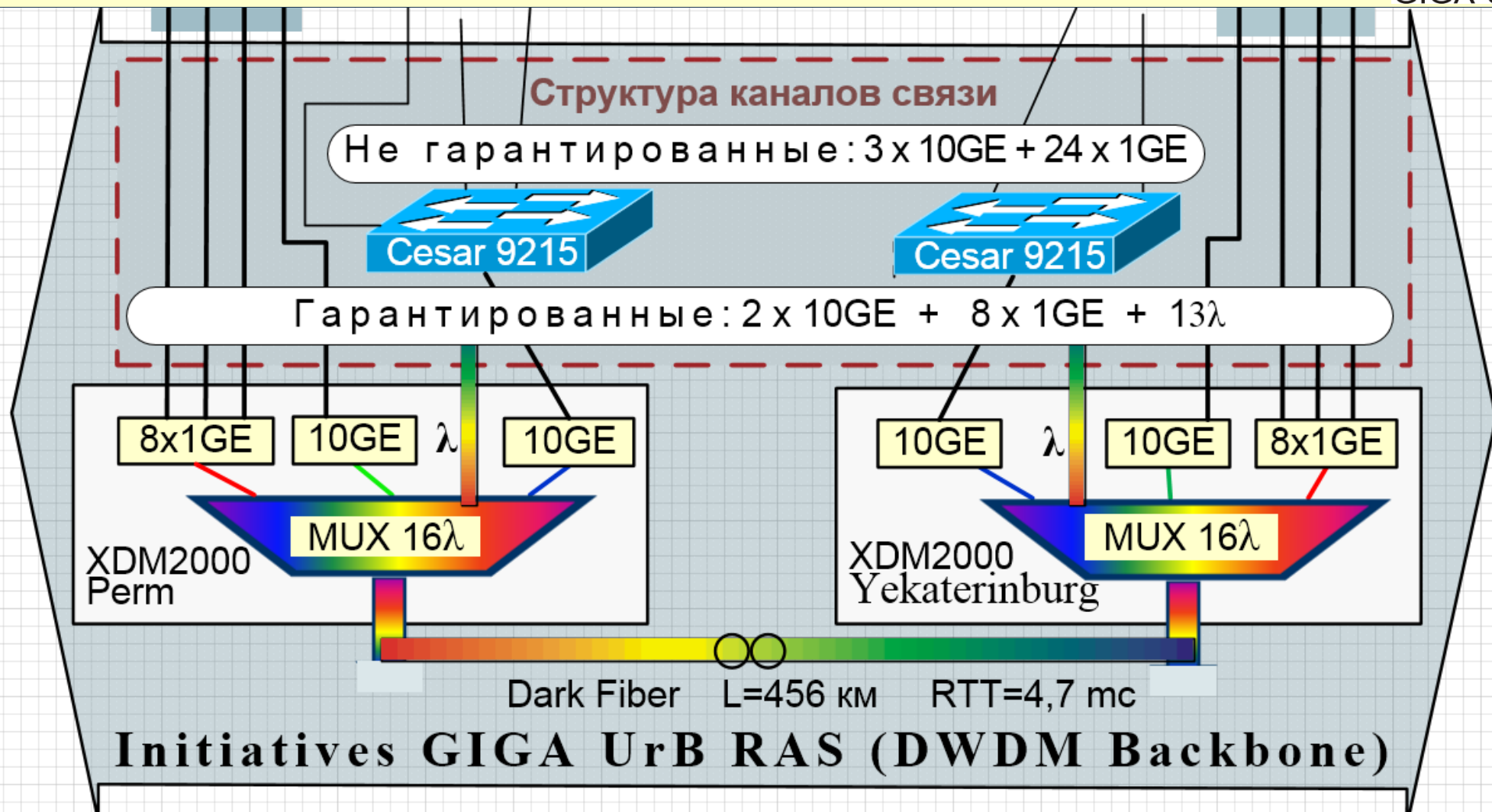
- **Судьба пропускной способности и динамика развертывания сети у институтов УрО РАН, а не у операторов связи**
- **Настраиваемый λ -коммуникационный сервис** на разнообразные потребности вместо одного арендуемого потока, приспособляемого для всех приложений / проектов
- **Фундамент создаваемой киберинфраструктуры региона**



Удачная организационно-территориальная структура:

- Монолитная структура УрО РАН, позволяет принимать неординарные и амбициозные решения
- Чем больше протяженность ВОЛС тем очевидней их преимущества

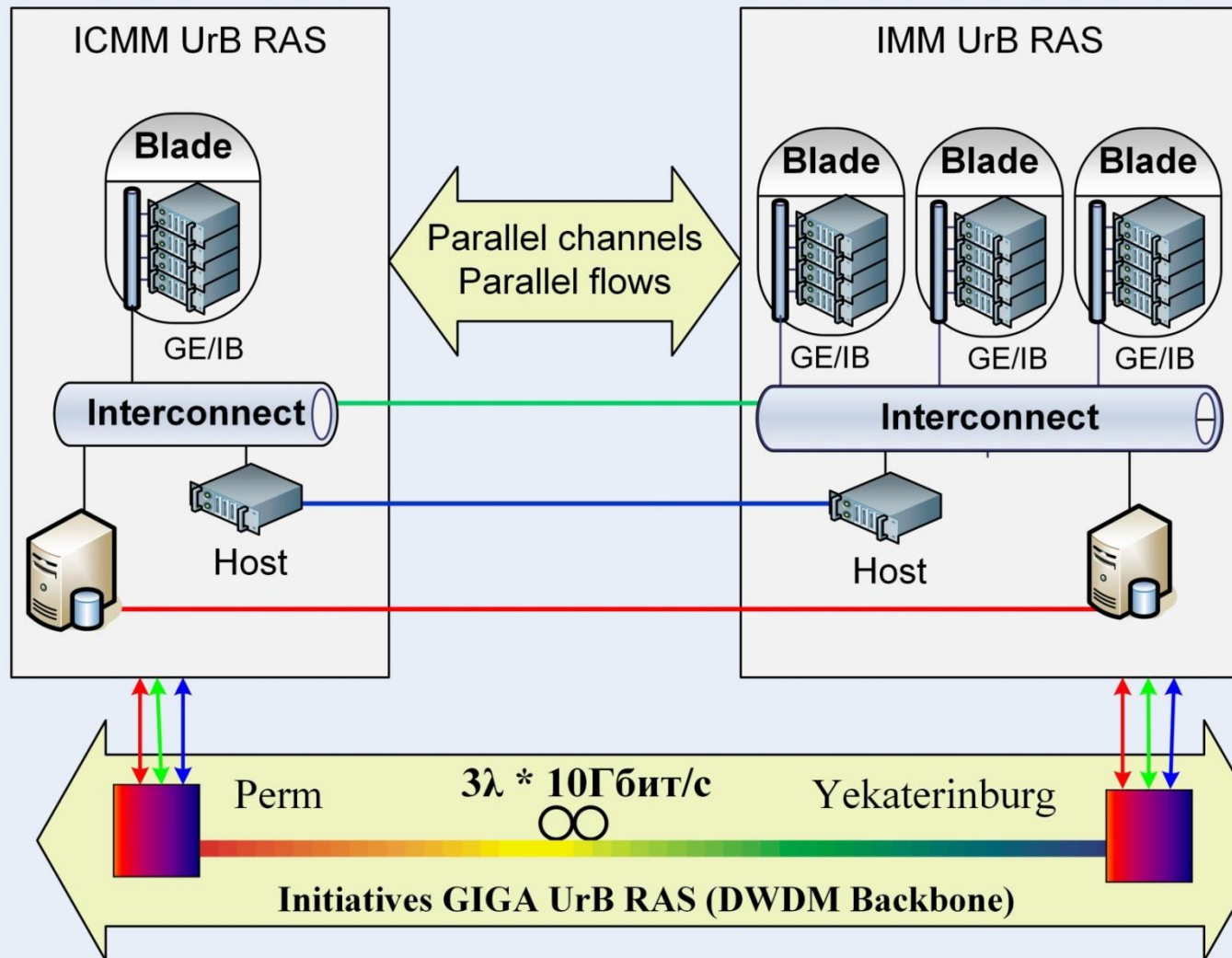
DWDM тракт Пермь-Екатеринбург (2012)



- λ услуга - фиксированный диапазон длин волн в DWDM магистральной (L1 OSI RM).
- FrameNet услуга – двухточечный 1-10 Гбит/с Ethernet канал связи: (L2 OSI RM)
 - гарантированная (DWDM) / негарантированная (Switch) магистральной
- PacketNet услуга – IP / присоединение к Public Интернет (L3 OSI RM)

Distributed Interconnect (2013)

Distributed Interconnect



Идея –

соединение по скоростной сети внутренних Interconnect кластеров и систем хранения данных.

Особенность

– параллелизм
лямбда каналов в DWDM и потоков данных в них

Цель – создание вычислительной инфраструктуры для исследователей, которые работают по различным аспектам параллельных, распределенных, сетевых и облачных вычислений.

Опыт: «Распределенный PIV» (2009-2011)

Киберинфраструктура территории

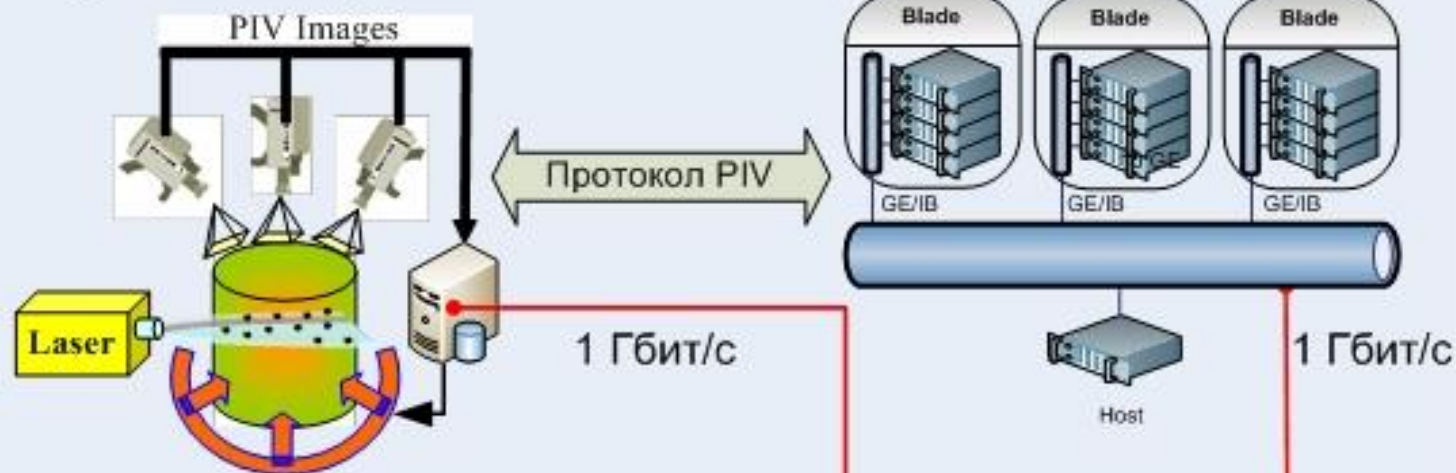
Проект Распределенный PIV

Пермь

Лаборатория физической гидродинамики
Экспериментальная установка PIV

Екатеринбург

Суперкомпьютерный центр
URAN, 160Tflops



Архангельск – Сыктывкар – Ижевск – Пермь – Екатеринбург

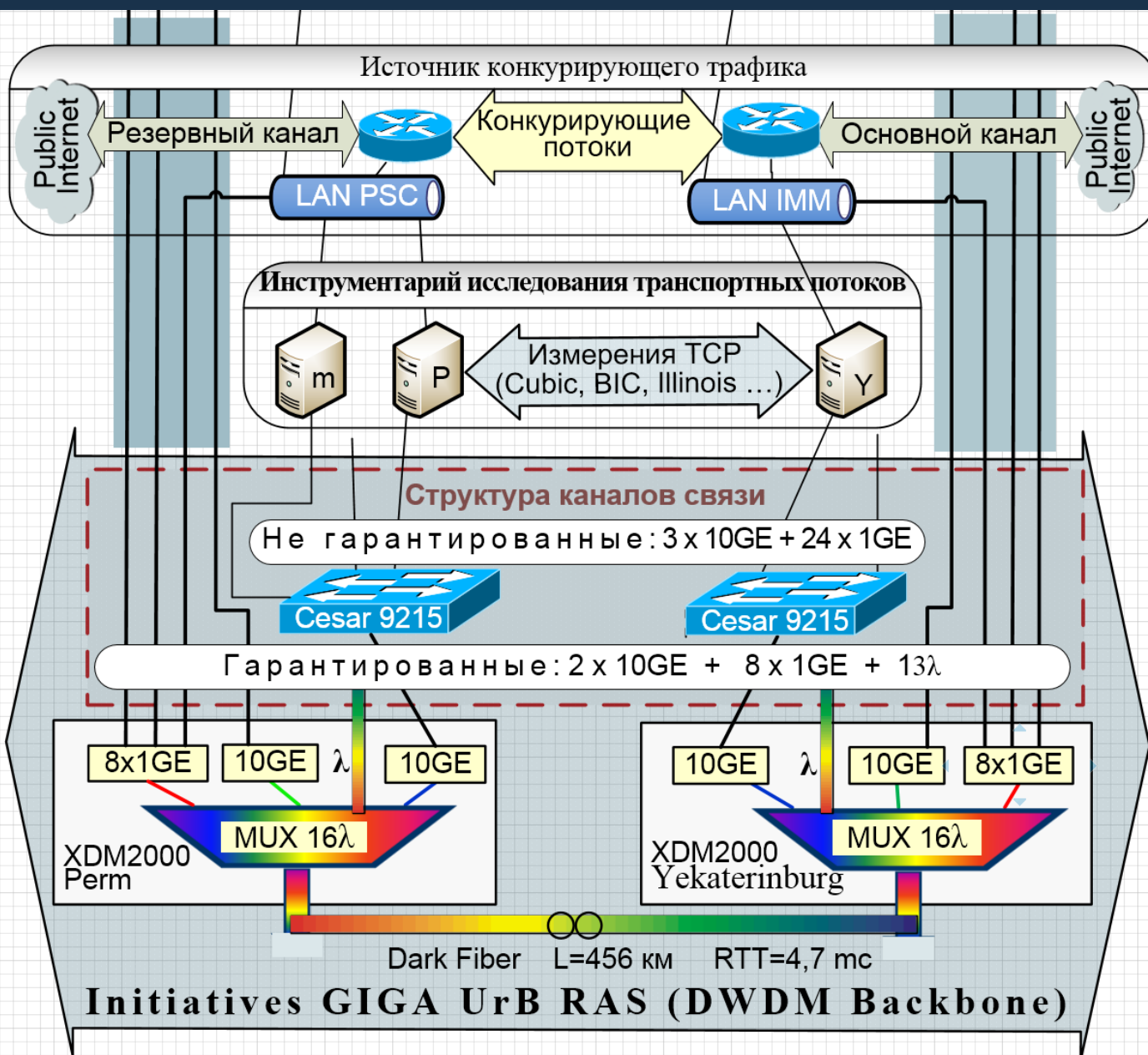
Initiatives GIGA UrB RAS (DWDM Backbone)

Цель – обработка в реальном времени потока экспериментальных данных

Идея – ввод интенсивного потока данных непосредственно в оперативную память вычислительных узлов суперкомпьютера

Успешность разработки новых парадигм распределенных вычислений возможно только **при тесном взаимодействии** архитекторов многопроцессорных систем и сетевых технологий, профессионалов по системному и параллельному программированию, ученых, экспериментаторов и постановщиков задач разных предметных областей

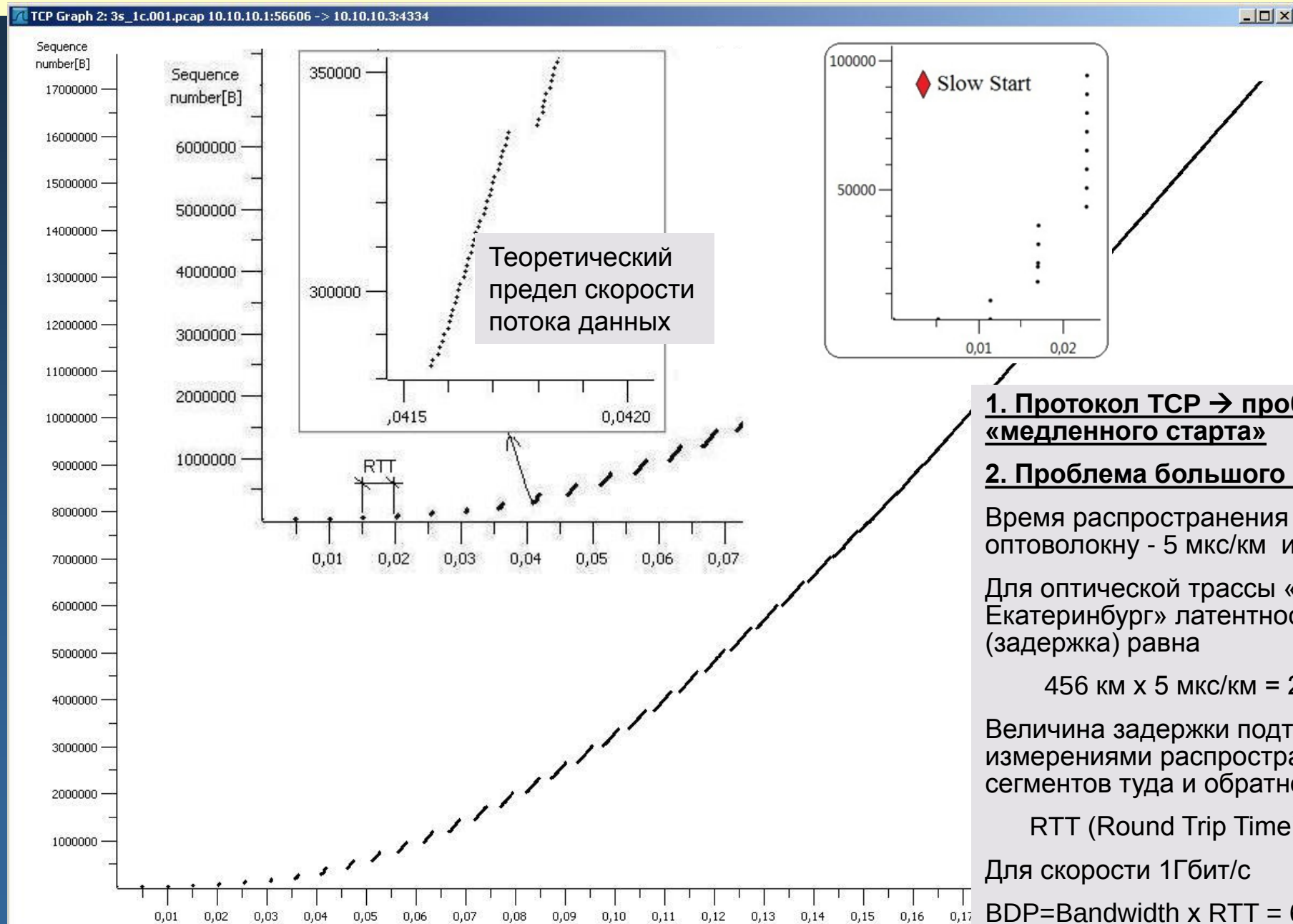
Distributed Interconnect – измерения



Одним из основных источников ухудшения совокупной производительности территориально распределенных высокоскоростных приложений является плохая end-to-end производительность протокола TCP

Доступные измерительные инструменты (ping, Traceroute, Tcpdump, Pchar, Iperf) не позволяют выявить первопричину плохой производительности TCP и диагностировать эффективность результатов настройки в TCP стратегии управления перегрузкой

Проблемы TCP по каналам с большим BDP



1. Протокол TCP → проблемы «медленного старта»

2. Проблема большого BDP

Время распространения сигнала по оптоволокну - 5 мкс/км и

Для оптической трассы «Пермь-Екатеринбург» латентность (задержка) равна

$$456 \text{ км} \times 5 \text{ мкс/км} = 2.3 \text{ мс}$$

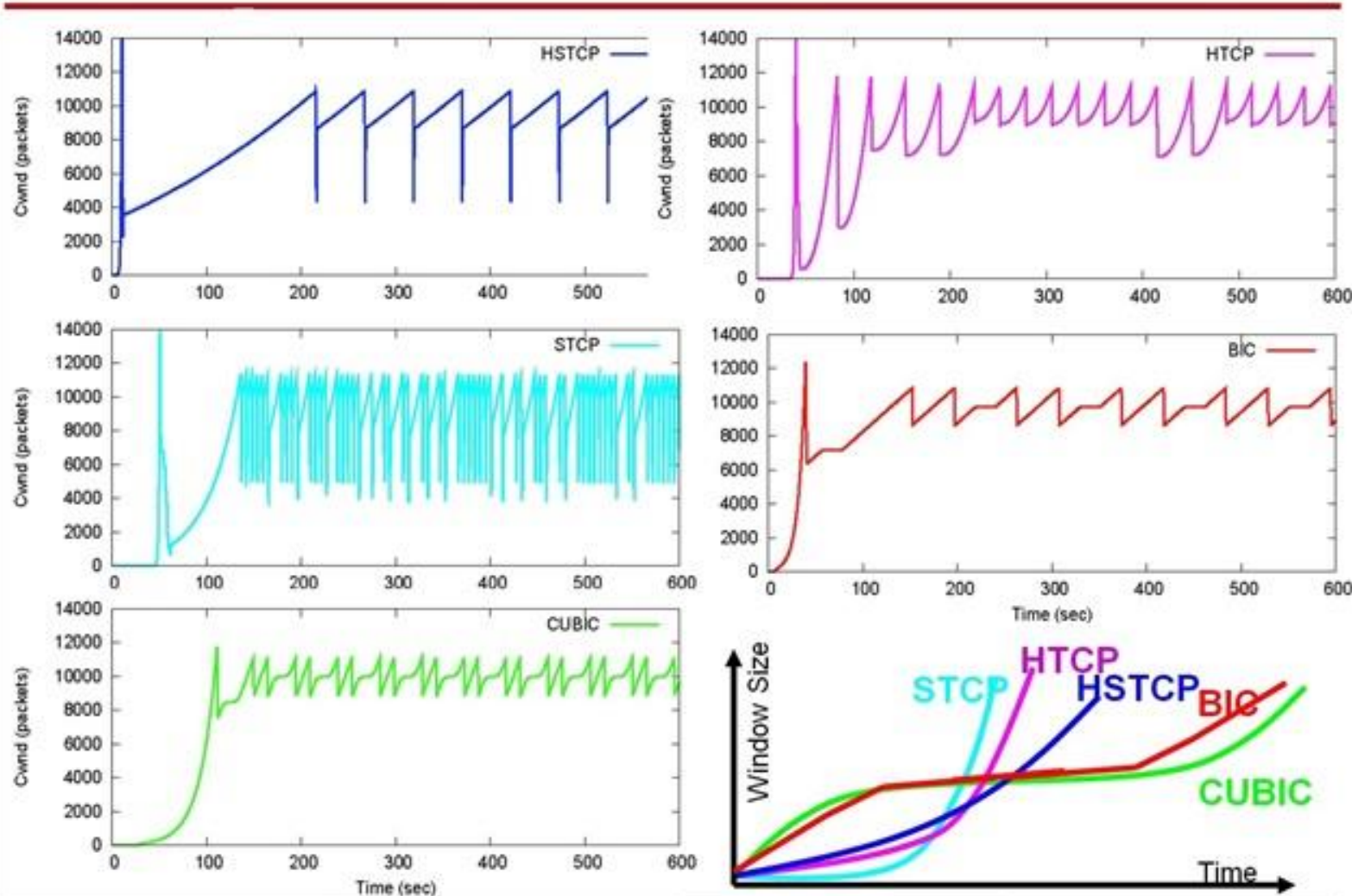
Величина задержки подтверждается измерениями распространения TCP сегментов туда и обратно

RTT (Round Trip Time) ~5 мс

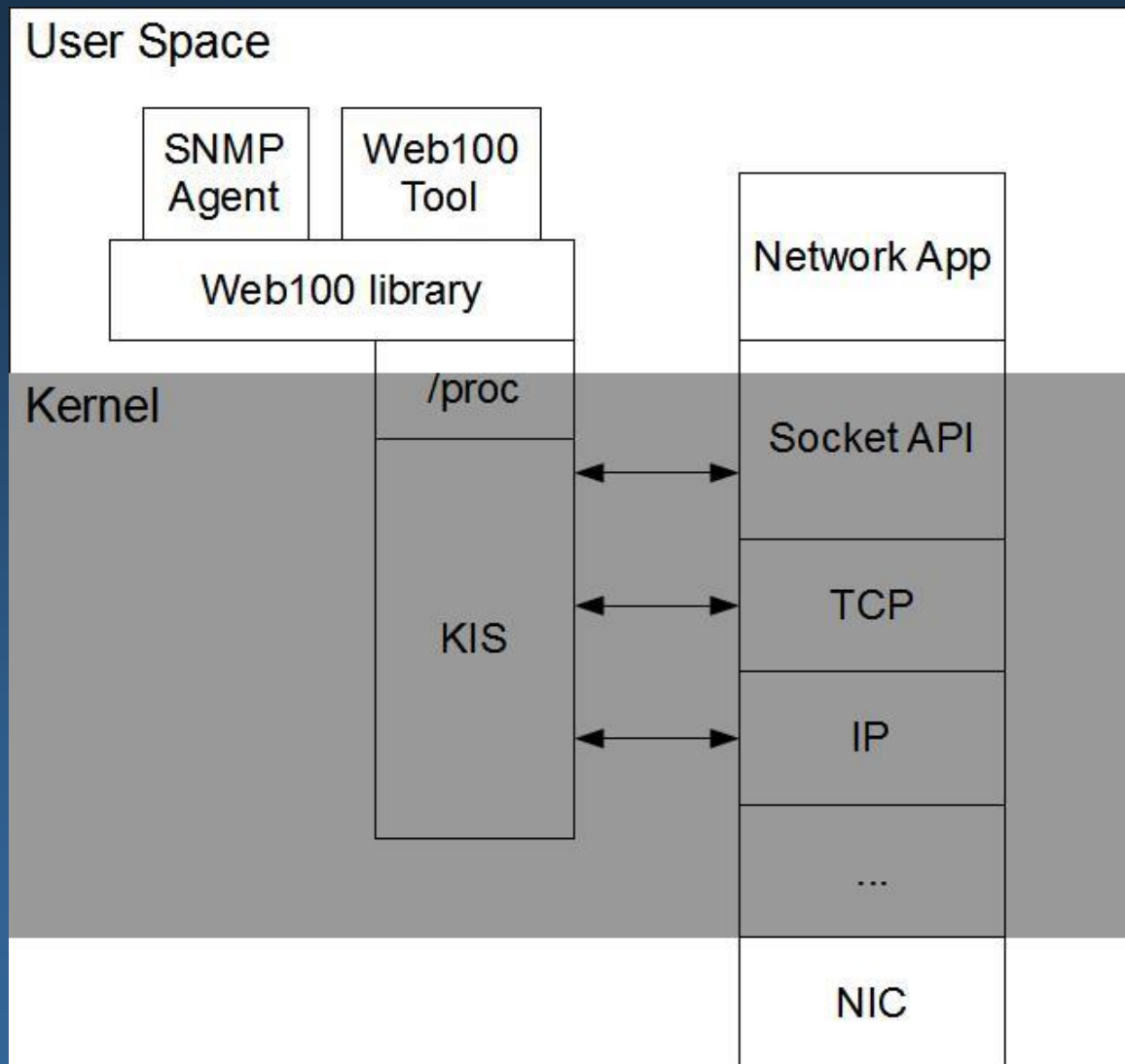
Для скорости 1Гбит/с

$$\text{BDP} = \text{Bandwidth} \times \text{RTT} = 625\,000 \text{ байт}$$

Example of Growth Functions



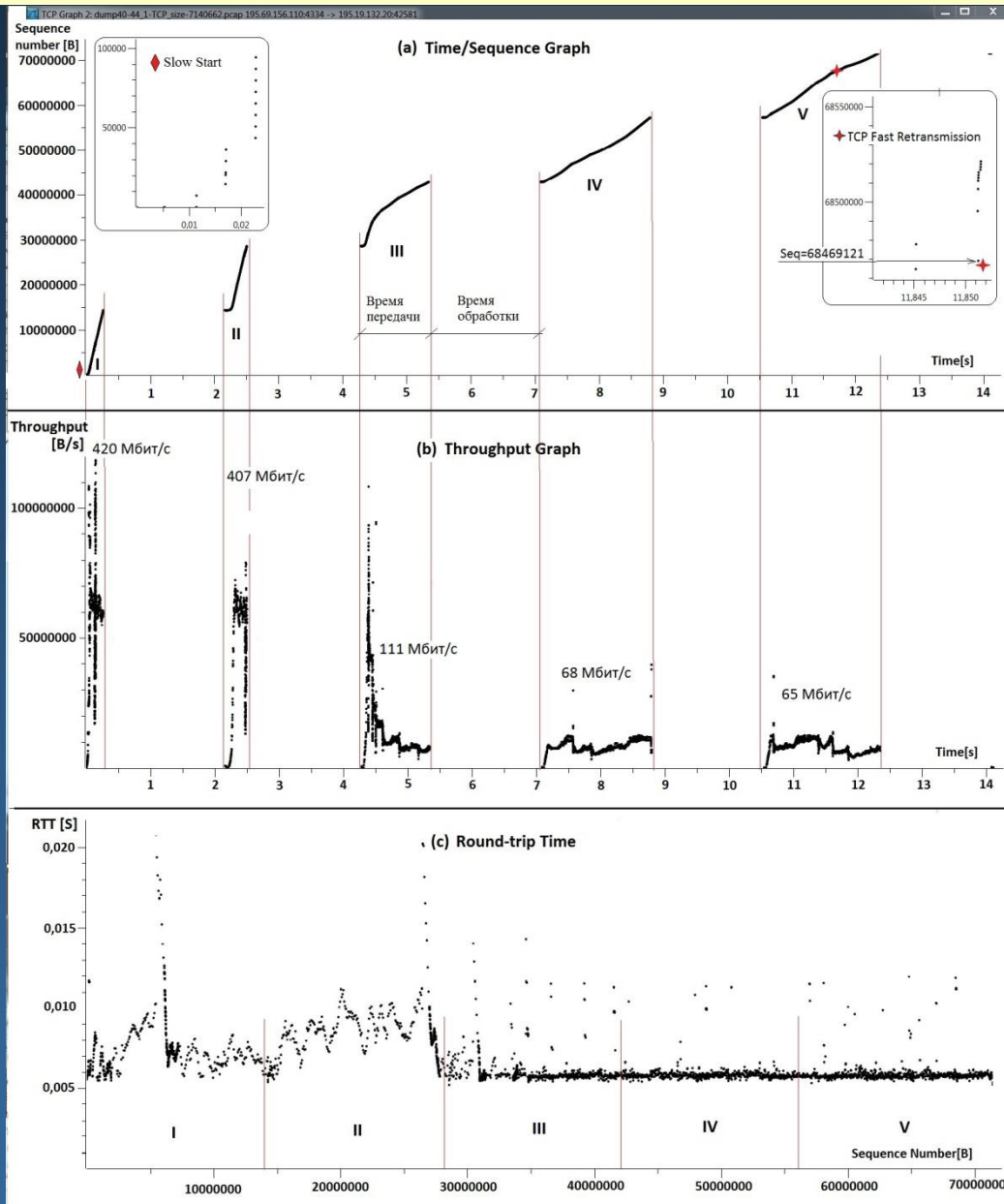
Source http://research.csc.ncsu.edu/netsrv/sites/default/files/stochastic_ordering_pfldnet_2007.pp



Цель Web10G — внедрение инструментария TCP в основную ветку ядра Linux.

Наличие базы управляющей информации *TCP Extended Statistics MIB* (RFC 4898)

Возможность создания собственных инструментов для взаимодействия с параметрами TCP.

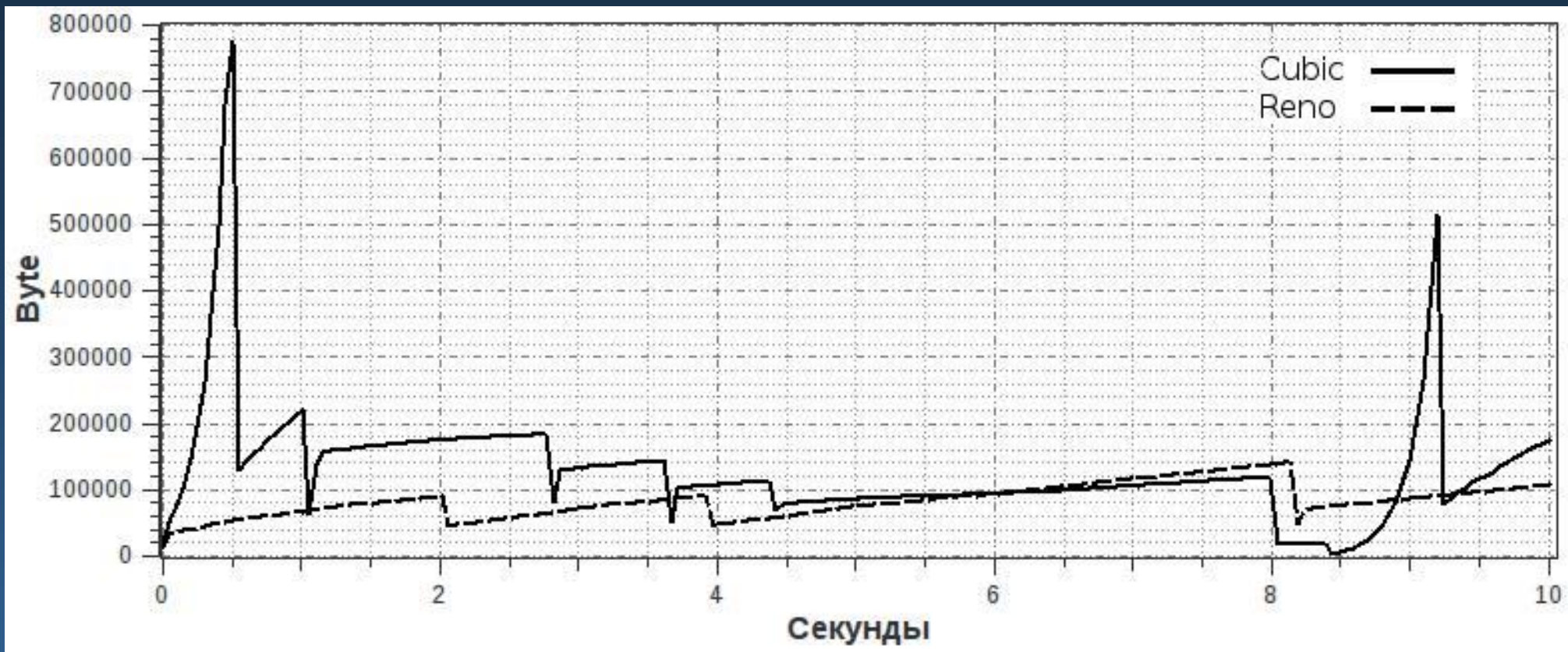


Цель измерений — определение возможной потери темпа передачи при переходе от одного кванта времени передачи к другому.

Снижение средней скорости передачи, вследствие дружелюбного отношения Cubic к конкурирующему трафику.

Подтверждается сохранение темпа передачи в промежутках между передачами данных.

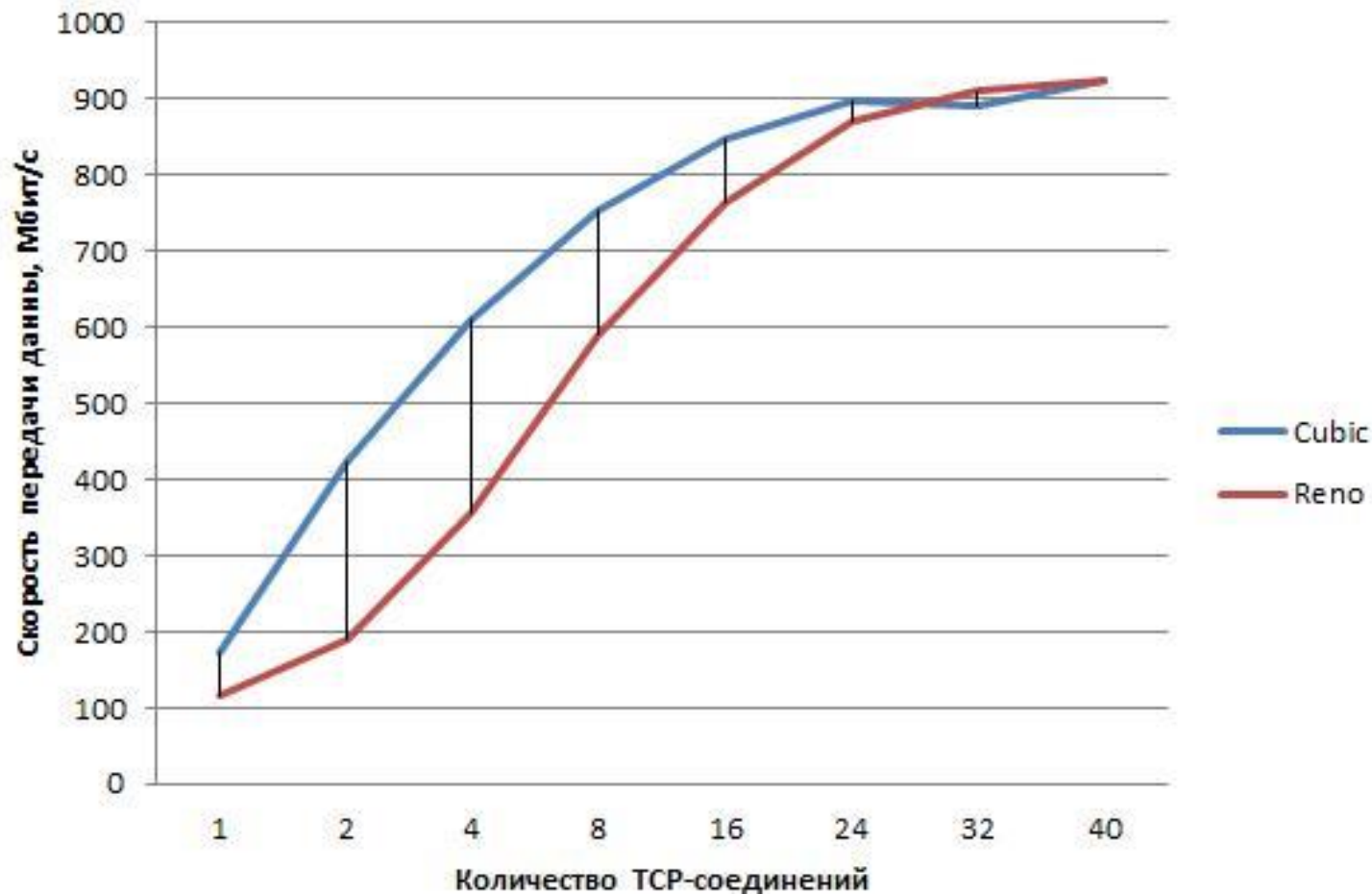
TCP - конкурирующие потоки



Недоиспользование пропускной способности канала связи одним потоком TCP.

Более агрессивное поведение стратегии Cubic, чем Reno.

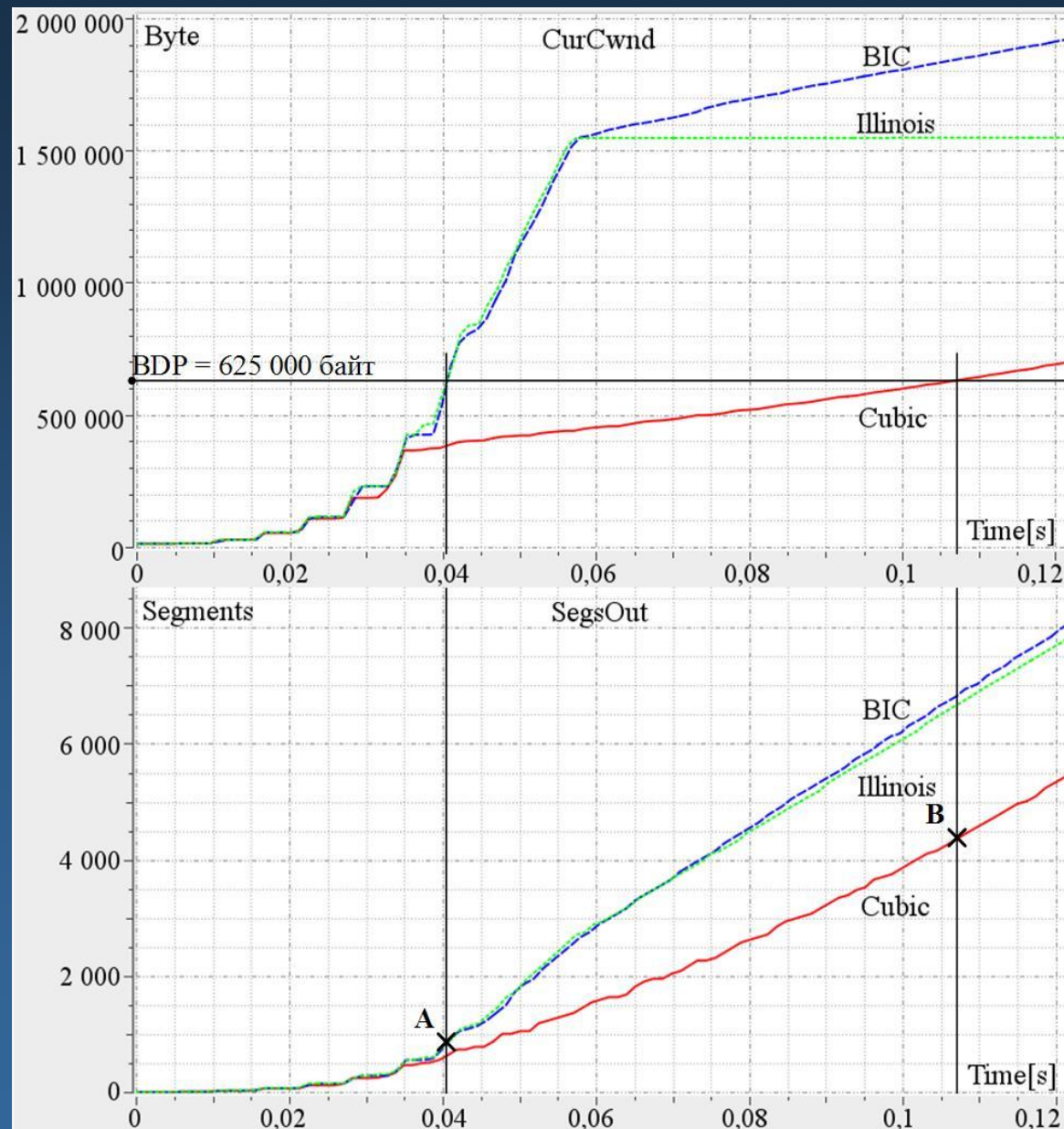
Параллельные ТСР потоки с конкурирующим трафиком



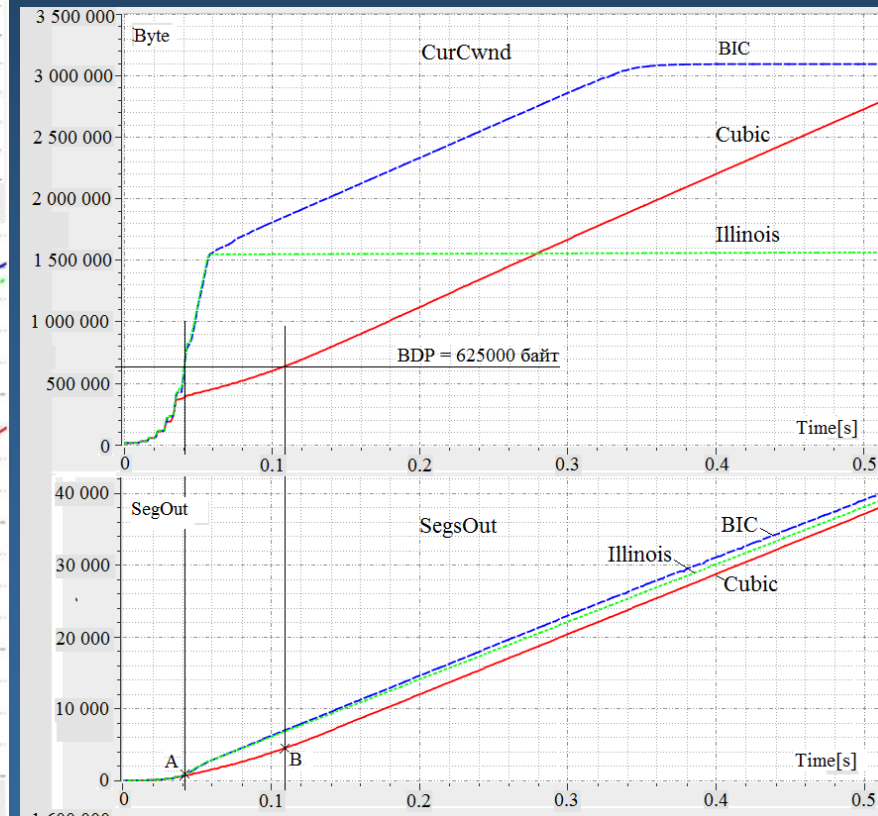
Цель – определение минимального числа ТСР потоков для утилизации пропускной способности канала связи.

Не рекомендуется использовать большое число потоков (более 8) из-за конкурирующего трафика.

ТСР потоки на выделенном канале



Различное начало перехода с процедуры медленного старта к основному режиму работы.



Выводы

Территориальная инфраструктура вычислителя существенно отличается от классических Grid, работающих через Public Internet.

Web10G инструментарий позволяет считывать переменные состояний TCP и объяснять истинные причины низкого темпа передачи.

Решение проблемы медленного старта – установка начального значения параметра *swnd* в оконечных системах равным BDP.

Число параллельных потоков не следует делать слишком большим (не более 8).

При передаче импульсного трафика в TCP соединении темп передачи в конце текущего и начале следующего интервалов передачи сохранятся.



Благодарю за внимание !