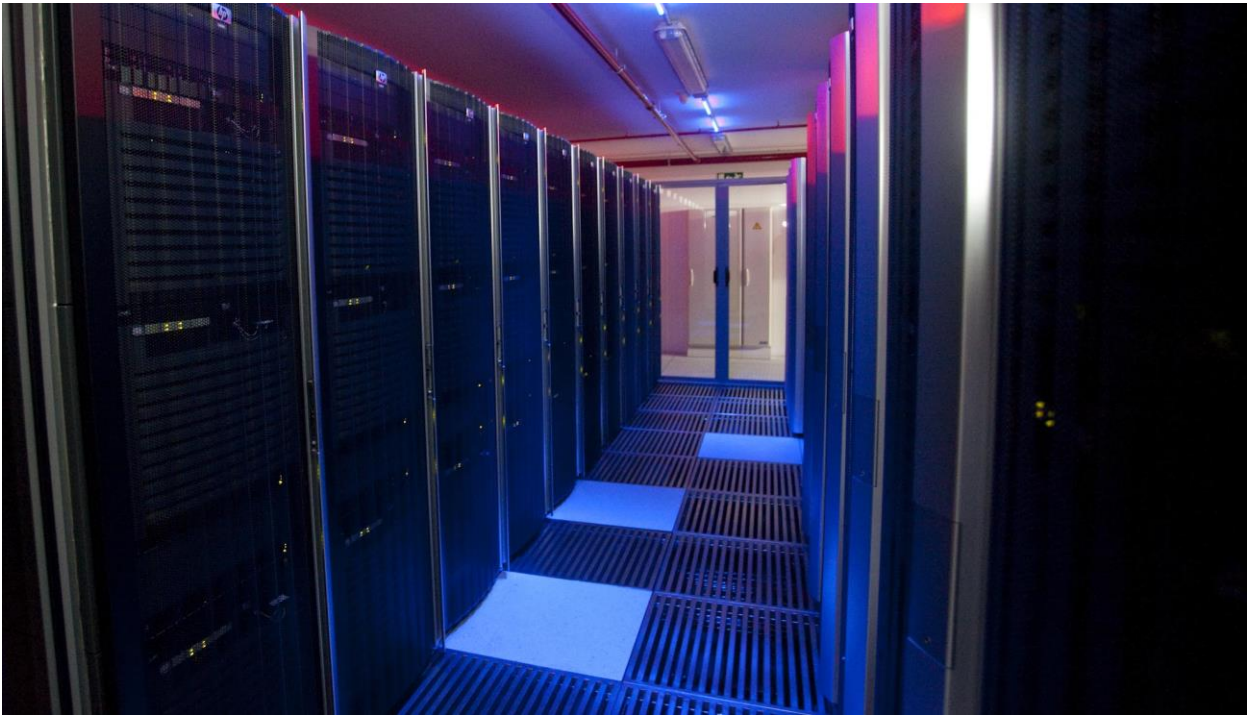
¿Son nuestros usuarios capaces de transferir datos de forma realmente rápida?

CESGA



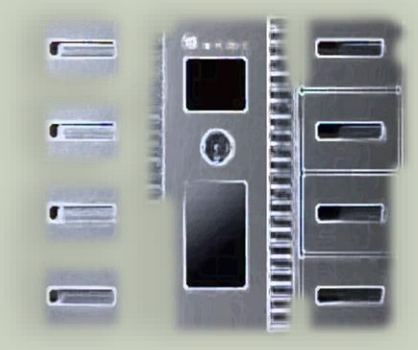
Infraestructuras

CESGA

Superordenador FT II
(7,712 cores)

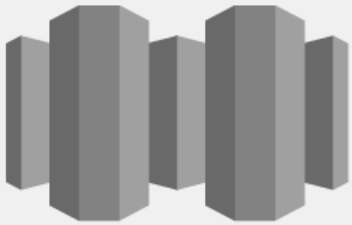
Superordenador SVG
(~3300 cores)

ALMACENAMIENTO
(1200 TB)



LIBRERÍA CINTAS
(2,200 TB)

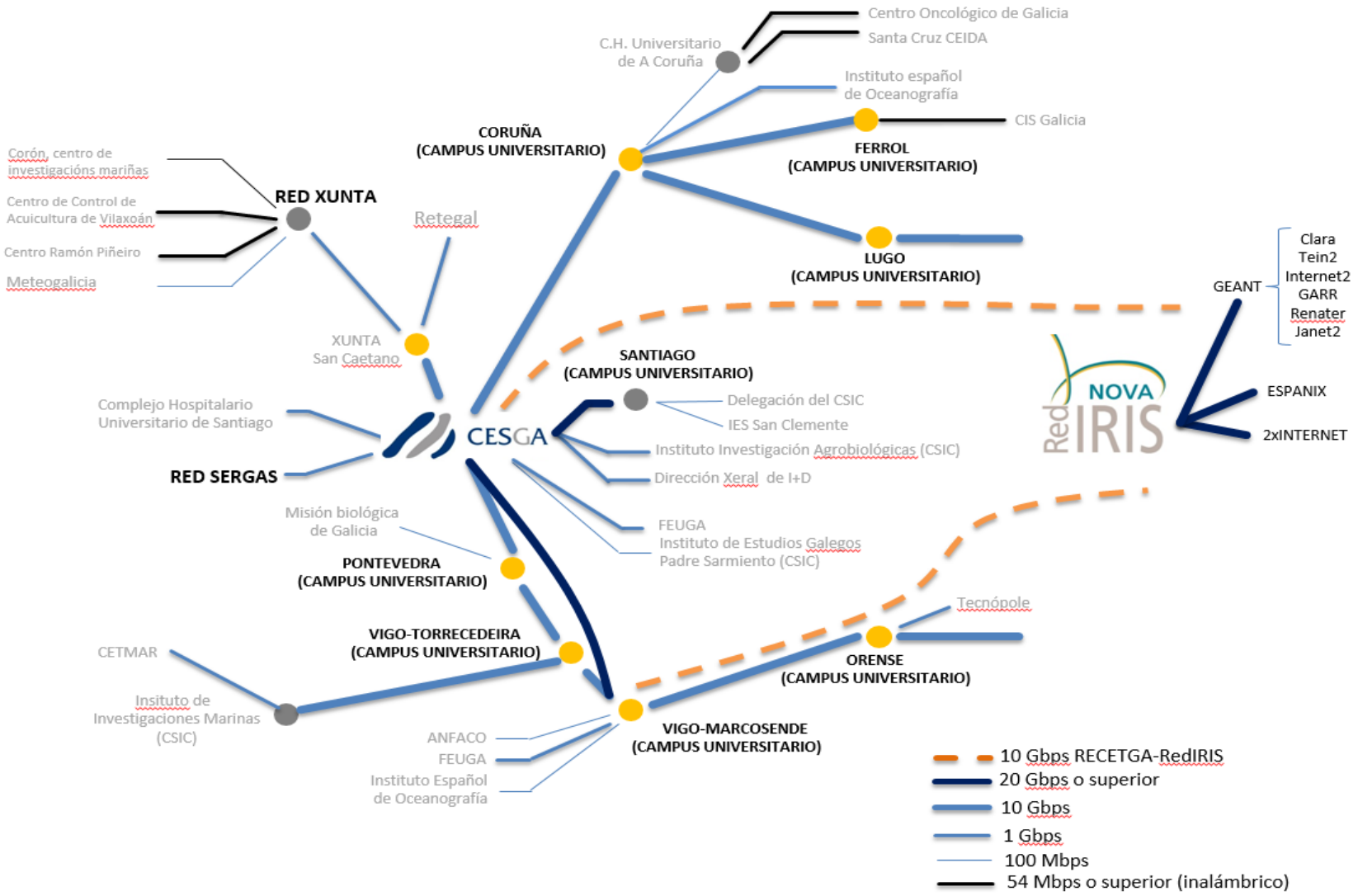
BIGDATA
(456 cores)



VISUALIZACIÓN
(80 cores)

CLOUD INDUSTRIA
(240 cores)

RECETGA


Centro Oncológico de Galicia
Santa Cruz CEIDA
C.H. Universitario de A Coruña
Instituto español de Oceanografía
CIS Galicia
CORUÑA (CAMPUS UNIVERSITARIO)
FERROL (CAMPUS UNIVERSITARIO)
Corón, centro de investigacións mariñas
Centro de Control de Acuicultura de Vilaxoán
Centro Ramón Piñeiro
Meteogalicia
RED XUNTA
Retegal
LUGO (CAMPUS UNIVERSITARIO)
GEANT
Clara
Tein2
Internet2
GARR
Renater
Janet2
XUNTA San Caetano
SANTIAGO (CAMPUS UNIVERSITARIO)
Complejo Hospitalario Universitario de Santiago
CESGA
Delegación del CSIC
IES San Clemente
Instituto Investigación Agrobiológicas (CSIC)
Dirección Xeral de I+D
RED SERGAS
NOVA
RedIRIS
ESPANIX
2xINTERNET
Misión biológica de Galicia
FEUGA
Instituto de Estudios Galegos Padre Sarmiento (CSIC)
PONTEVEDRA (CAMPUS UNIVERSITARIO)
Tecnópole
VIGO-TORRECEDEIRA (CAMPUS UNIVERSITARIO)
CETMAR
ORENSE (CAMPUS UNIVERSITARIO)
Insituto de Investigaciones Marinas (CSIC)
ANFACO
FEUGA
Instituto Español de Oceanografía
VIGO-MARCOSENDE (CAMPUS UNIVERSITARIO)
10 Gbps RECETGA-RedIRIS
20 Gbps o superior
10 Gbps
1 Gbps
100 Mbps
54 Mbps o superior (inalámbrico)

Usuarios

- Ámbito Gallego
  - Universidades gallegas
  - Centros de enseñanza Secundaria
  - Instituto Español de Oceanografía
  - Centros Xunta con I+D+i
  - Centros Tecnológicos
  - Unidades investigación hospitales
- Ámbito Nacional
  - Centros CSIC
  - RES
  - Proyectos
- Ámbito Europeo
  - Proyectos
  - Otras colaboraciones

Casos

- Envío de datasets para cómputo u obtención de resultados
  - Transferencia BSC, Proyecto MSCO4SC…
- Extensión de datacenter
  - ITER-CESGA
- Otros movimientos de datos bulk
  - Transferencias grabaciones JJTT

¿Cuánto puede un elefante correr?

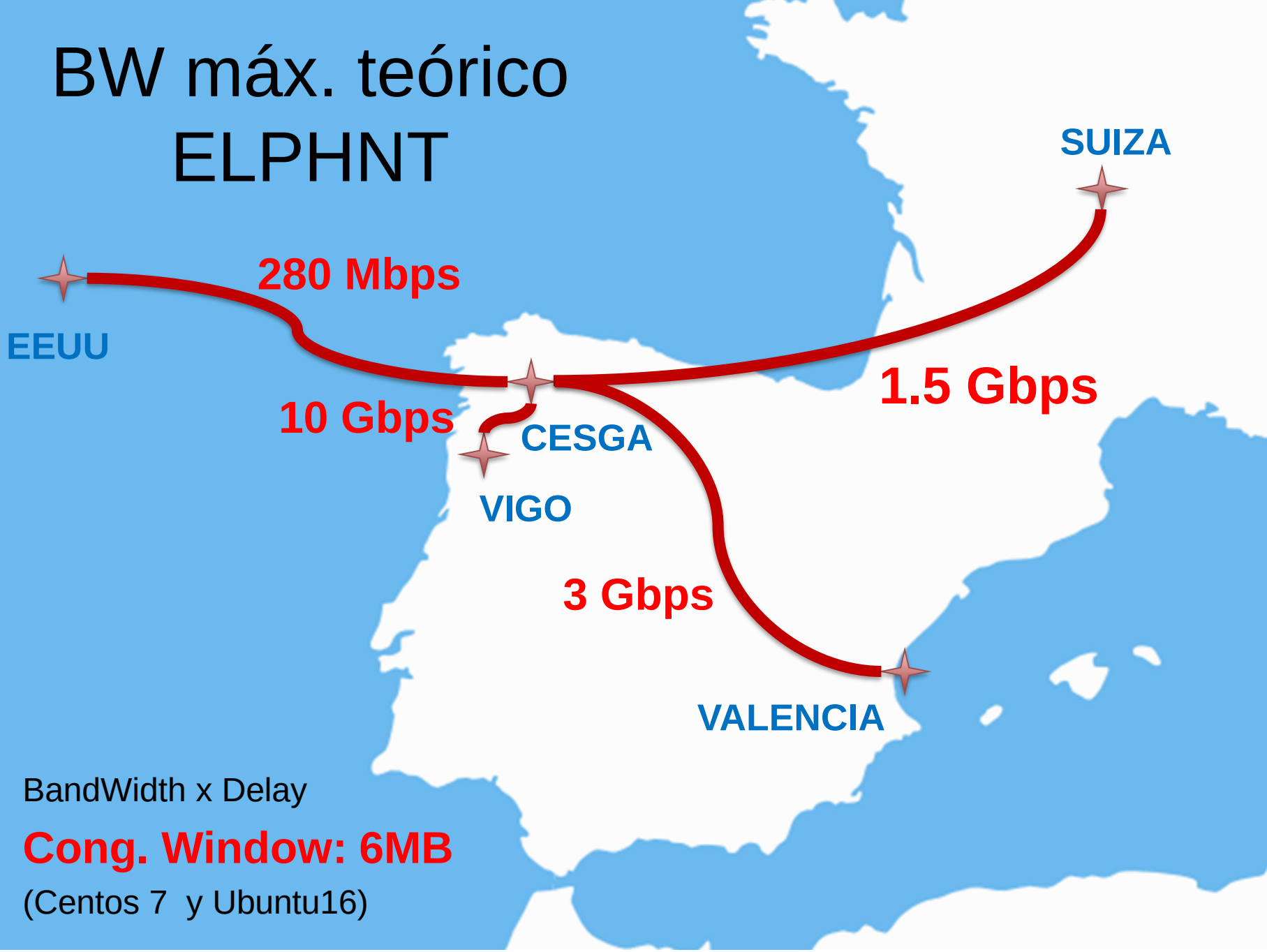
Credit: John Lund Stone Getty Images

Ratones vs Elefantes

Bulk
Deep Queue ok
Drops bad

VoIP
Deep Queue bad
Drops OK

BW máx. teórico ELPHNT



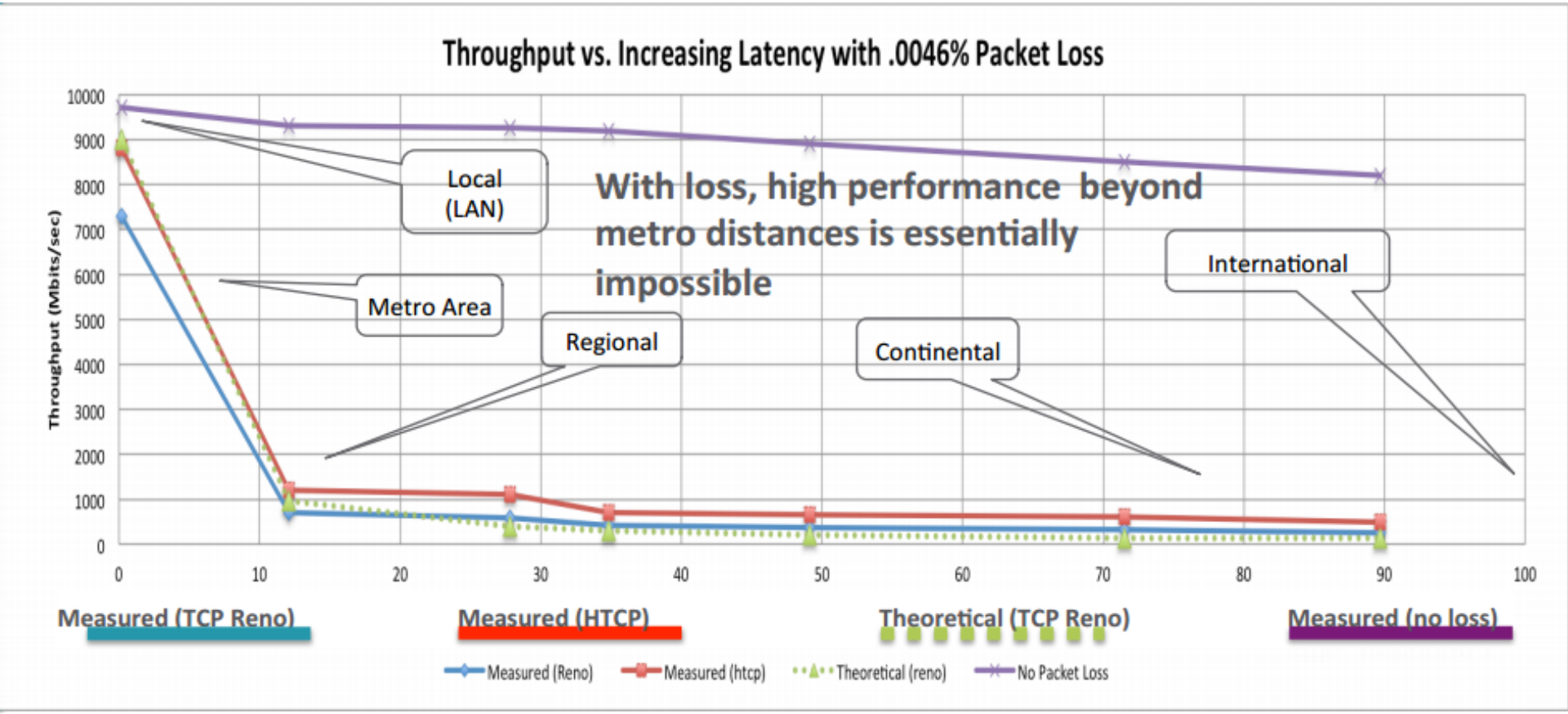
BandWidth x Delay

Cong. Window: 6MB

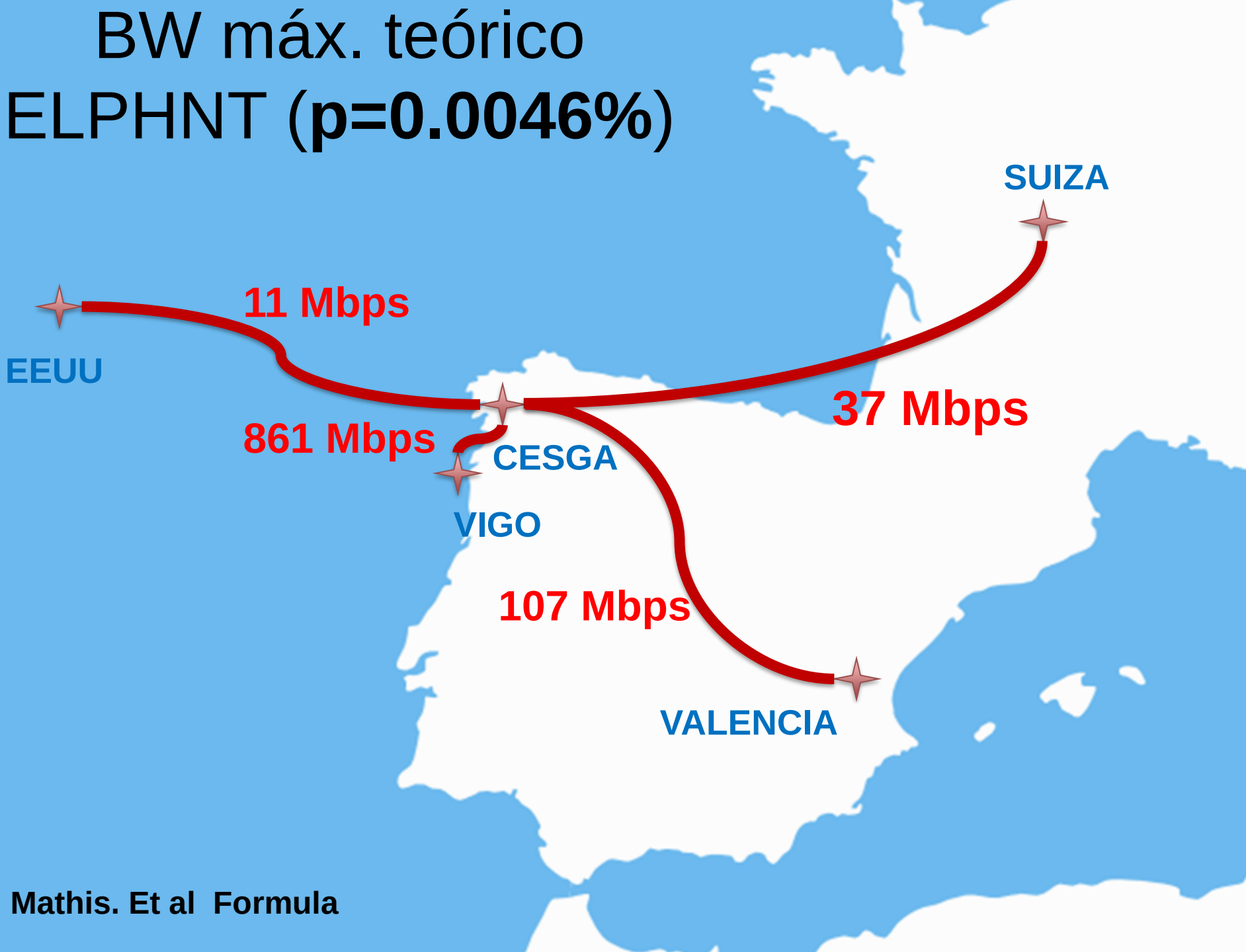
(Centos 7 y Ubuntu16)

¿Y con pérdida de pkts?

Caso ESNET

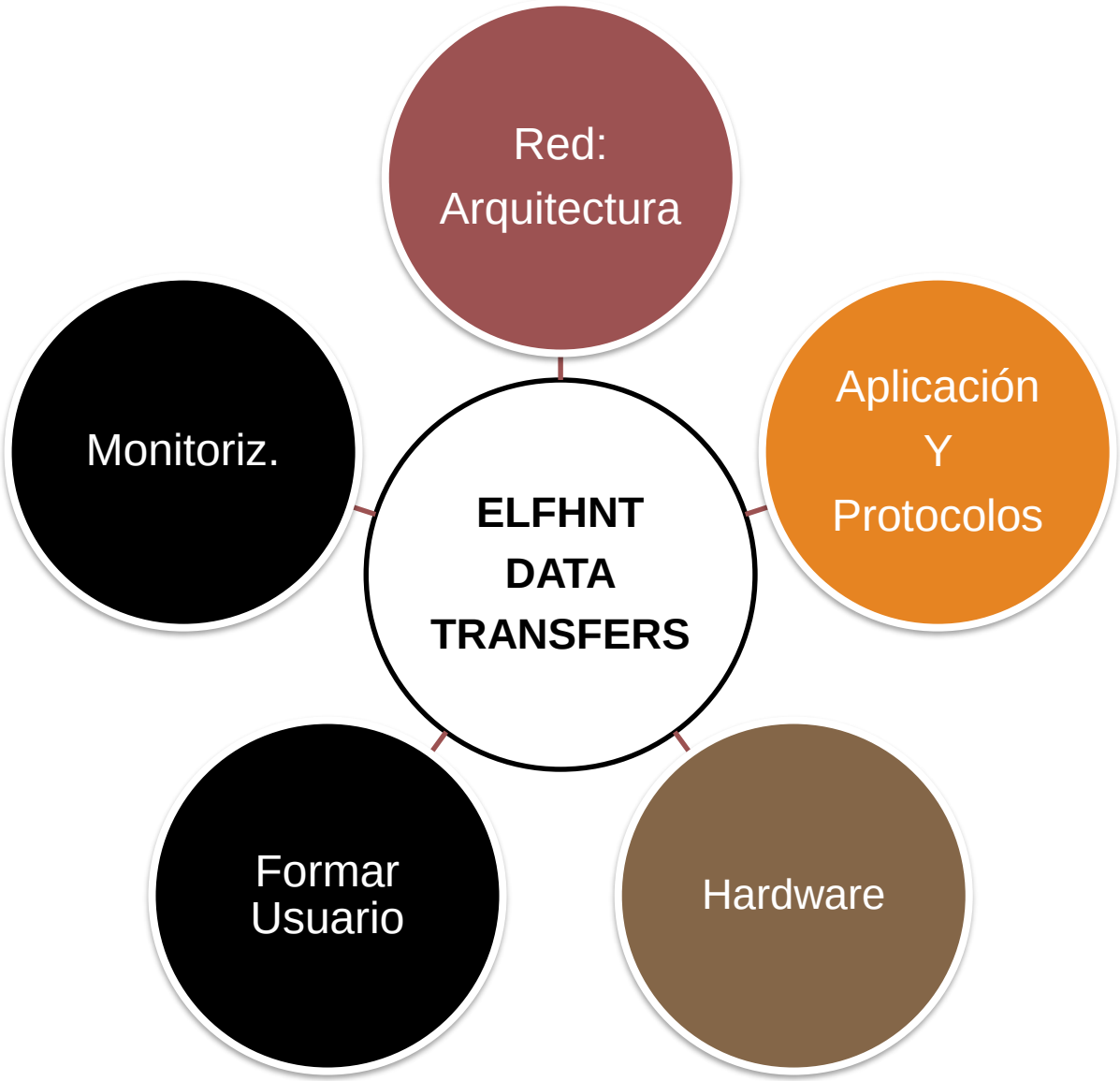


BW máx. teórico ELPHNT (**p=0.0046%**)

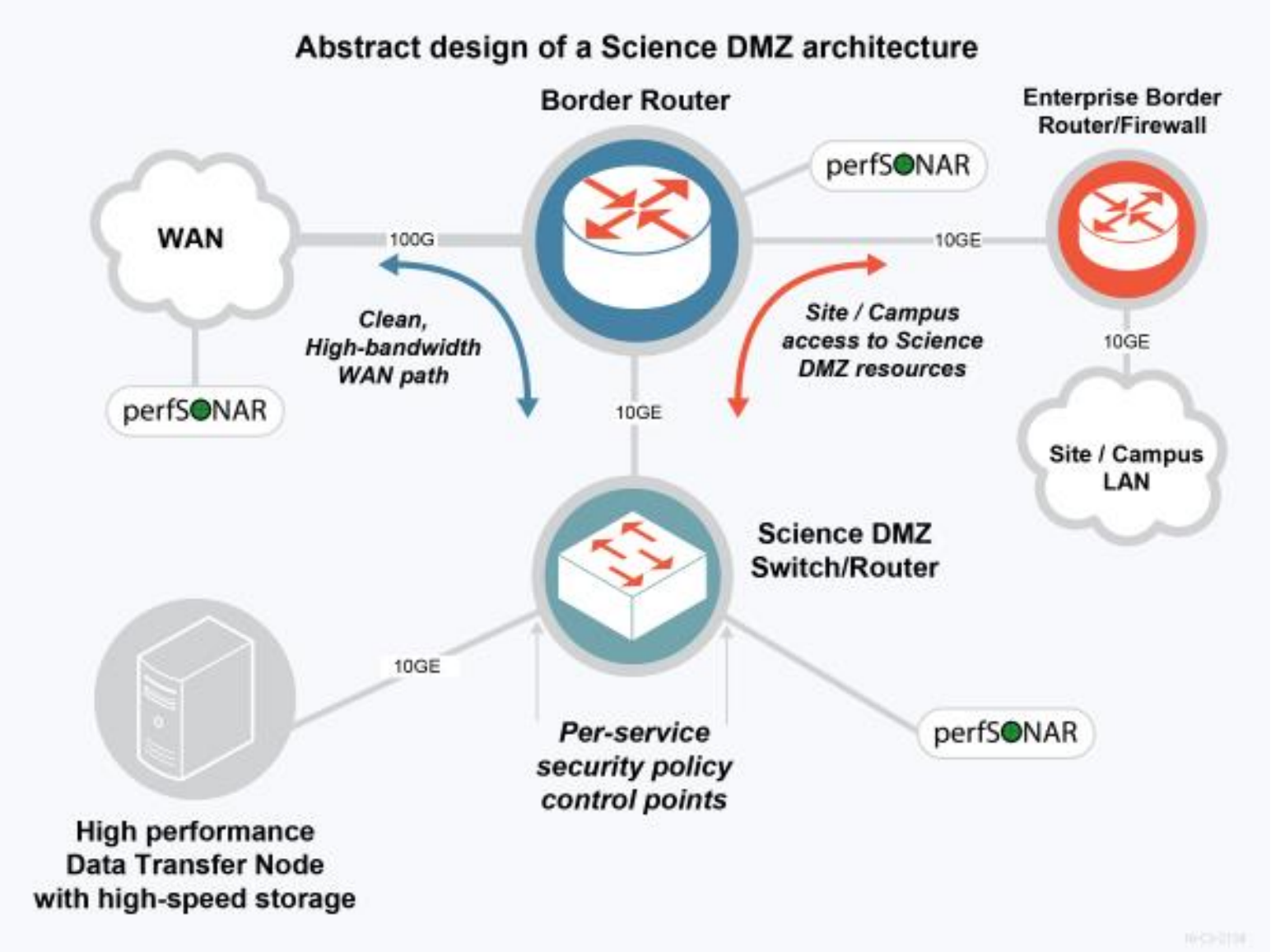


Mathis. Et al Formula

Mejora de las transferencias ELPHNT



Arquitectura de red – Science DMZ



Arquitectura de red – Science DMZ

- Construir la red para soportar Bulk Data Transf
  - No conectar en cualquier punto
  - Evitar atravesar la infraestructura Empresarial si es possible
  - Conectar los recursos de para BDT lo más cerca possible del router frontera
- Configurar routers y switches para buffering adecuado.
  - Vigilar drops y errors (microbursts y sobresuscripción)
  - Monitorizar periódicamente el rendimiento
- Creación de DTNs

ESnet's Science DMZ Breaks Down Barriers, Speeds up Science

SEPTEMBER 17, 2015

Contact: Jon Bashor, jbashor@lbl.gov, +1 510 486 5849

From individual universities around the country to a consortium of research institutions stretching the length of the west coast, networking teams are deploying an infrastructure architecture known as the Science DMZ to help researchers make productive use of ever-increasing data flows.

SCIENCE DMZ SPECIAL REPORT

CASE STUDY: UNIVERSITY OF UTAH »

CASE STUDY: GENERAL ATOMICS »

The Science DMZ traces its name to an element of network security architecture. In a security context, a DMZ or "demilitarized zone" is a portion of a site network which is specifically dedicated to external-facing services (such as web and email

Aplicaciones y hardware

Aplicaciones

- Número de flujos: scp, rsync, GridFTP, parallel rsync
- Limites búferes: scp, sftp, hpn-ssh
- Limitada por cores: tinc, openvpn
- Uso de cifrados: none, aes256-gcm,…

Servidor

- CPU
- Disco
- NICs y estrategias de packet pacing

Hardware de red

- Firewalls, conformadores de tráfico, etc
- Switches, Routers
- ¿Qué especificaciones tenemos de los fabricantes?

Hardware

https://people.ucsc.edu/~warner/buffer.html

Information here is by **rumor, innuendo and extrapolation**. Manufacturers rarely put info on packet buffers in their data sheets. There are some **summary thoughts**

The *buffer size* question discussed in 2012 on the nanog list and is reproduced. **Buffer requirements** for long RTT networks is less well understood than you might hope. IETF tests for burst management are not as well developed as bandwidth tests. There is a draft that hints at progress. Packet pacing can improve buffer effectiveness by making TCP less bursty.

Incast is a buffer exhaustion phenomena that is one consequence of running out of packet memory.

Shared memory means that the hardware permits buffers to be used by any port that needs them. An intel white paper compares shared memory with other architectures. In a shared memory design it is not possible to let ALL the memory go to queued packets. There would be no room for new arrivals which would lead to head of line blocking. The other major option for a switch fabric is a **crossbar matrix**.

Buffer queue depth monitoring cannot be done directly with SNMP MIB-II polling of the current occupancy of the buffer. Even if a buffer depth SNMP poll object existed it would not possible to interrogate it on a time scale short enough to catch microbursts. A burst that would fill a 4 MByte buffer would completely drain in 3.2 mS at 10 Gb/s. You could hope for indirect evidence of buffer exhaustion by monitoring packet drops. Bursts too short to cause drops can nonetheless be long enough to affect performance. Direct queue monitoring can thus add valuable information.

Some switches have multiple switch ICs that each manage their own memory pool. Examples are the Brocade FCX648S and the Cisco 3750-48. Memory from one IC can be shared among the ports in that IC's group but cannot be loaned out to ports controlled by other switch chips. Here we are interested in queue resources that can be claimed by a single flow for burst absorption -- not the total RAM in the system.

Tolly (tolly.com) occasionally reports on the ability of switches to sustain **microbursts** in his reports on data center switches. These measurements relate directly to output port buffering. See esp the IBM G8264 below.

Max buffer queue depth requires that all packet memory can be put into a single queue. QoS schemes divide buffer resources among defined queues. As such, I am not interested in the QoS descriptions and these are even less reliable than the rest of this doc.

Model	Port Type	RX Queue	TX Queue	Total Buffer	RX Buffer	TX Buffer
Trident+ Shared Memory						
Accton 5652	48 SFP+ and 4 QSFP+	8Q		9 MB		5? MB
Edge-corE 5600-52X	48 SFP+ and 4 QSFP+	8Q		9 MB		5? MB
Juniper QFX3500	48 SFP+ and 4 QSFP+	8Q		9 MB		5 MB
Arista 7050S-64	48 SFP+ and 4 QSFP	8Q		9 MB/switch		5 MB
Dell 8132F & 4032F	24 SFP+ and 2 x QSFP+	8Q		9 MB		
Dell 8164F & 4064F	48 SFP+ and 4 x QSFP+	8Q		9 MB		
Pica8 P-3920	48 SFP+ and 4 x QSFP+			9 MB		
Penguin 4804x	48 SFP+ and 4 x QSFP+			9 MB		
Cisco Nexus 3064X	48 SFP+ and 4 QSFP+	12Q		9 MB		5 MB
Supermicro SSE-X3348T	48 10GTw-Pr and 4 QSFP+	8Q		9 MB		

Protocolos

TCP

- Tuning
- Variantes

Transfrencia fiable UDP

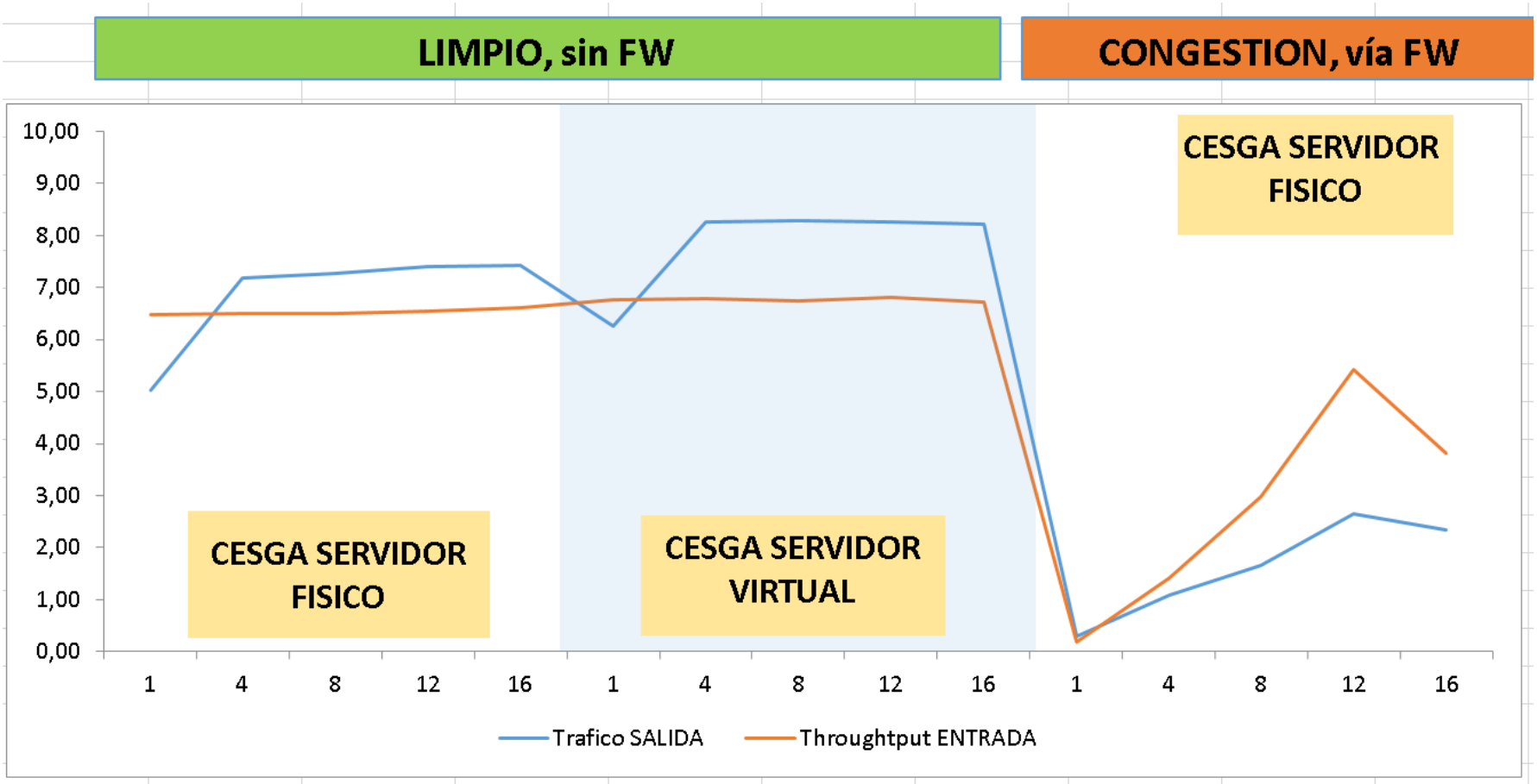
- UDT (GridFTP)
- FASP (Aspera)
- Quic (Google)

Monitorizar

- Eventos:
  - Errores y drops
  - **Medidas periódicas**

- Herramientas:
  - Perfsonar
  - Mastema

Medir y verificar



Medidas para la mejora

Ya existe:
- DMZ Ciencia
- Conocimiento límites

Medidas en curso

Medición de rendimiento:
- DMZ de ciencia y empresarial
- Físico/virtual

Y actuaciones:
- Reencaminamiento flujos
- Apertura tickets fabricantes

Mejora monitorización:
- Detectar mejor drops
- Medidas periódicas rendimiento

Seguimiento próximo de casos:
- Usuarios de servicios
- Caso JJTT2017 y sonda TELTEK

Need 4 speed

- Fuentes de información
  - ESNET Fasterdata Knowledge Base
    - An Expert Guide for End-to-End Performance Tuning, Tools and Techniques
    - Objective: Enabling the highest levels of performance for the Department of Energy (DOE) scientific community
  - EduPERT knowledge base
    - Federated PERT that links the independent PERTs (GÉANT PERT, National, Local and Project PERTs) with a portfolio of central services to aid them in their network investigations.


The End...

Referencias

1. HPN-SSH / Pittsburg Supercomputing Center
   https://www.psc.edu/hpn-ssh
1. SSH Performance / Allan Jude, ScaleEngine Inc. allanjude@freebsd.org
   http://www.allanjude.com/bsd/AsiaBSDCon2017_-_SSH_Performance.pdf

The ISP Column. Faster. Geoff Huston
   http://www.potaroo.net/ispcol/2005-06/faster.html