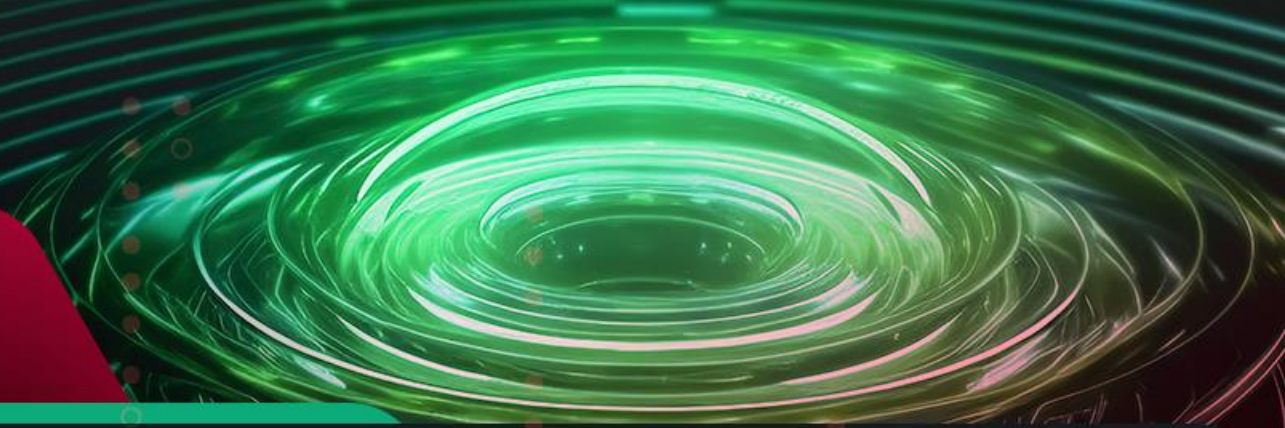
Bicsi®
ENDORSED EVENT

ICT SUMMIT MEXICO 2024

El futuro de ICT - Inteligencia Artificial, Convergencia y Sustentabilidad



24 DE ABRIL

Hotel Galería Plaza, San Jerónimo, Ciudad de Mexico.

ORGANIZA:

LATAM RED

www.latamred.com

Latam Red S.A. latamred

Diseño de Cableado en Centros de Datos para nuevas aplicaciones en la era de la IA

Pedro Lerma, RCDD/DCDC/NTS/OSP/RTPM
Enterprise Sales Manager, Mexico
Corning Optical Communications

Agenda

Descripción de las redes a diseñar en los Data Centers

Consideraciones para las redes de Front end/producción

Características de la IA y las redes de Back-end

Resumen & Q/A

Optical Internetworking Forum (OIF)

Es un organismo que promueve el Desarrollo y la Implementación de soluciones para la interoperabilidad de las redes y Servicios a través de acuerdos para productos ópticos de networking, elementos de procesamiento de redes y diferentes componentes. Estos acuerdos se basan en los requerimientos de Desarrollo y coperación entre usuarios finales, proveedores de Servicio, fabricantes y proveedores de tecnología; alineados con los estándares internacionales o aumentados si es necesario.

Esto se logra mediante la participación de los miembros de la industria que trabajan juntos para desarrollar especificaciones para:

- Interfaces de elementos de red externos
- Interfaces de software internas para los elementos de red.
- Interfaces de componentes de hardware internas para elementos de red

La OIF crea puntos de referencia, realiza pruebas de interoperabilidad a nivel mundial, genera conciencia en el mercado y promueve la educación sobre tecnologías, servicios y soluciones. El OIF proporciona comentarios a las organizaciones de normalización de todo el mundo para ayudar a lograr un conjunto de soluciones interoperables e implementables

Evolución en la velocidad de los Switches en el DC Pasado, Presente y Futuro

1RU Frontplate Density (32Ports):

Puertos 3.2T
- Co-empaquetamiento óptico (224G XSR/DD)
- OSFP-XD (16x224G VSR)

Puertos 1.6T
- Co-empaquetamiento óptico (112G XSR)
- QSFP-XD (16x112G VSR)

Puertos 800G
- Co-empaquetamiento óptico (56G XSR)
- OSFP (8x112G VSR)

Puertos 400G
- QSFP-DD (8x56G VSR)
- OSFP (8x56G VSR)

Puertos 100G

640G 40nm

1.28T 40nm

3.2T 28nm

6.4T 16nm

12.8T 16nm

256x50G

25.6T 7nm

256x100G
512x50G

Breakout a 200G/400G FR/DR

51.2T 5nm

256x200G
512x100G

102.4T 3nm

256x400G
512x200G

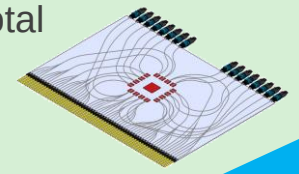
204.8T 2nm

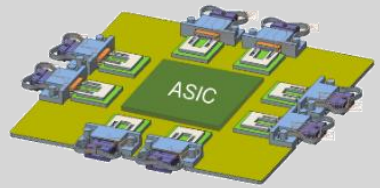
512x400G
1024x200G

409.6T <2nm

1024x 400G

Oportunidad para una solución óptica total





oportunidad para co-empaquetamiento

400G Aumento & Comercialización

2028 2030

Ruta a 400G

400G SerDes
o SerDesLess

★ OIF-224G-XSR 200G SerDes
o SerDesLess

★ OIF-112G-XSR 100G SerDes

★ OIF-112G-VSR 100G SerDes

★ OIF-56G-VSR 56G SerDes

OIF

2010 2012 2014 2016 2018 2020 2022 2024 2026

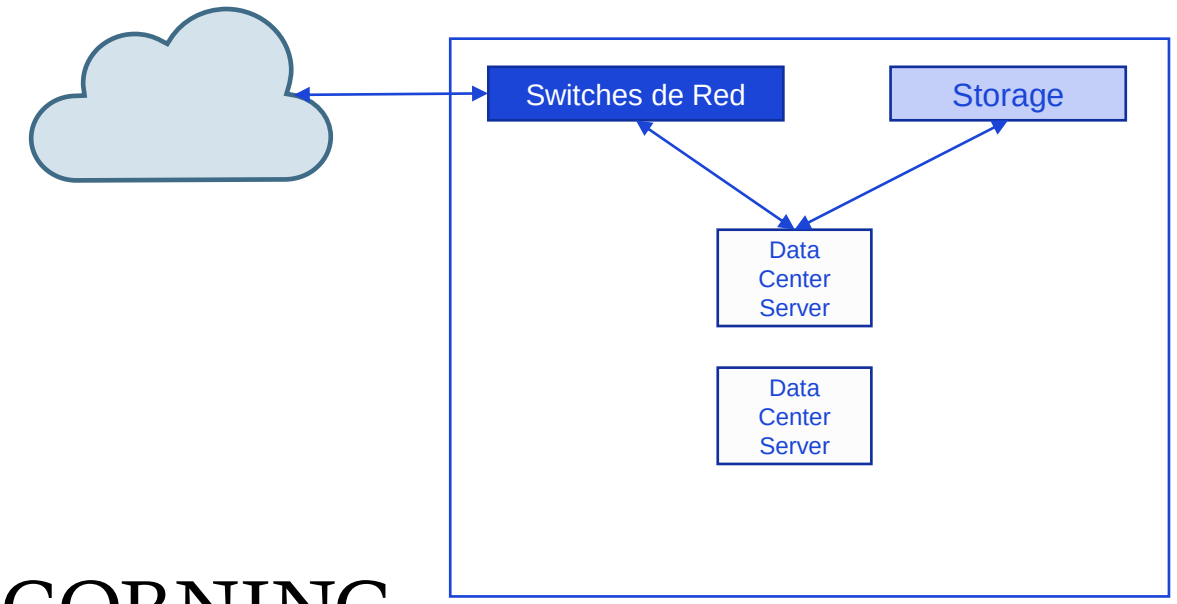
CORNING

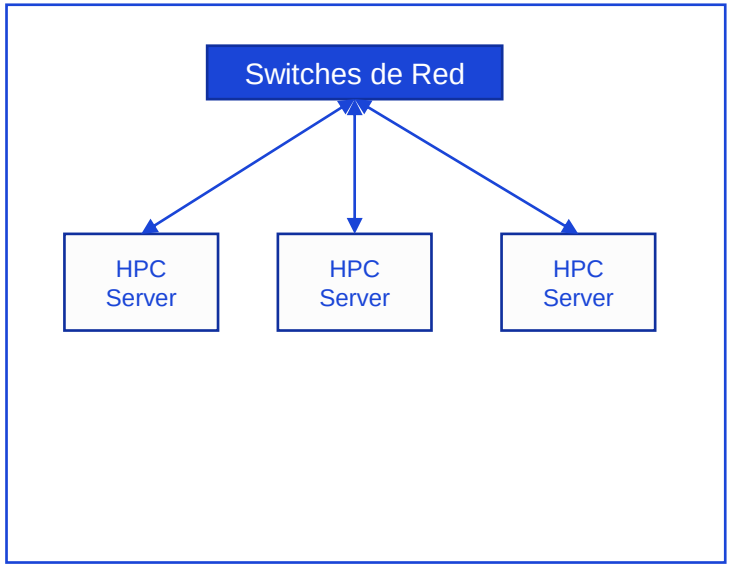
Bicsi ENDORSED EVENT

Redes dentro del Data Center

Un centro de datos moderno se compone de diferentes redes

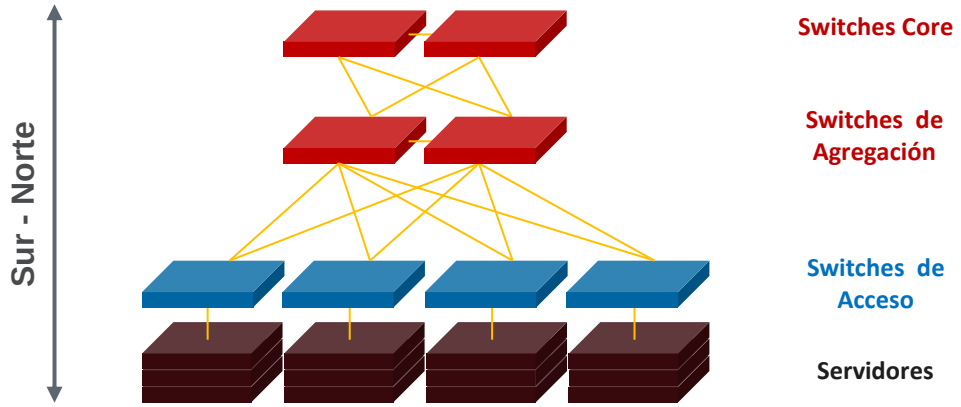
- Redes de Front End/Producción: Servers del Data Center corriendo aplicaciones, redes de Almacenamiento
  - Protocólos Ethernet & Fiber Channel
  - El Tráfico entra y sale de las instalaciones
- Back End Network (IA/ML)
  - Servidores de Computo de Alto Desempeño (HPC)
  - Protocolos de red de InfiniBand
  - El Tráfico permanece dentro de las instalaciones (server to server)





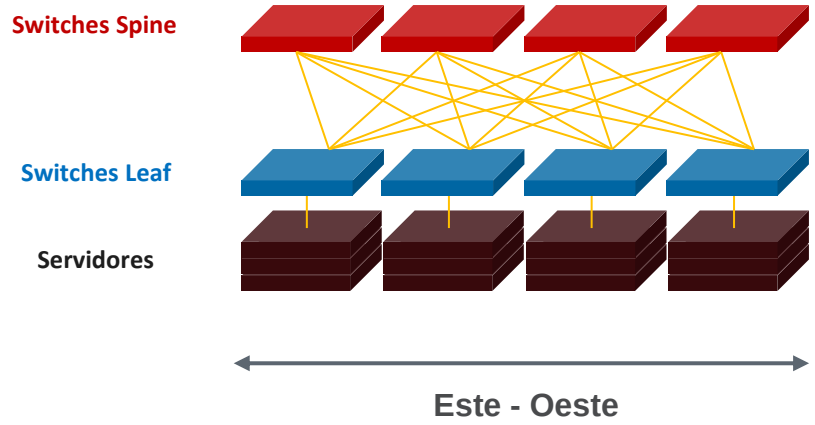
Evolución de la Topología de la Red

Arquitectura Tradicional de Red "3-Tier" de árbol



- Los Datos viajan jerarquicamente (Flujo de datos de Sur a Norte)
- La ruta más larga agregará laterncia y creará cuellos de botella
- Solo permite una escalabilidad limitada
- La Redundancia es limitada

New "2-Tier" Architecture Spine-Leaf



- Los Datos viaja de Este a Oeste
- Cada Switch del Leaf está conectado a cada Switch del Spine, lo que significa que la información cruzará la misma cantidad de dispositivos, minimizando los cuellos de botella
- Proporciona una Escalabilidad/Expansión más sencilla
- Proporciona redundancia, si una de las Spines llegara a fallar, tendría un impacto menor en el DC

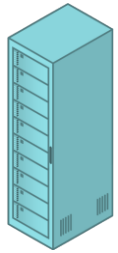
Elementos de la Red del Front End/Producción dentro de un Data Center

Requisitos para la interconexión del servidor de Front End

Tipos de Servicios

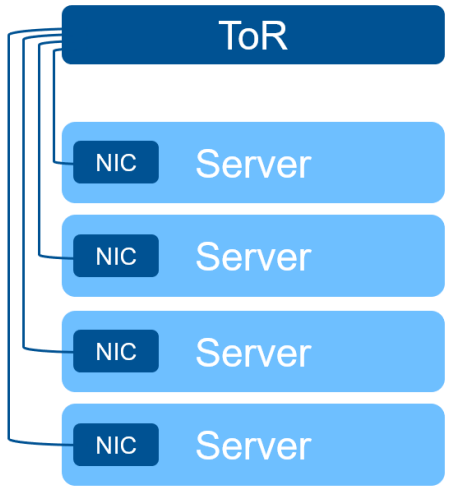
Emoresa Grande

Empresa Pequeña



- **Servidor Web.** Aloja sitios web, incluidos todos los datos del sitio web
- **Servidor de E-Mail.** Facilita el envío y recepción de correo electrónico
- **Servidor de Base de Datos**. Almacena datos en el backend y los recupera de las computadoras en el frontend mediante consultas como SQL
- **X – Servidor de Servicios**. Google Maps, HPC, etc

Diseño General



Consideraciones de Diseño

- Requisito para conectar servidores y elementos de la red de almacenamiento a través de una serie de switches
- Conectividad de Servidor/Almacenamiento al switch de red realizada a través de DAC, AOC o SFP
- Tasa de Transferencia de Datos de 1 Gbe – 25 Gbe
- La sensibilidad de la Aplicación a la latencia determinará la cantidad de sobresuscripción aplicada a los switches de red
- El Procesamiento es manejado principalmente por la CPU dentro del servidor

Atributos del Cableado Estructurado

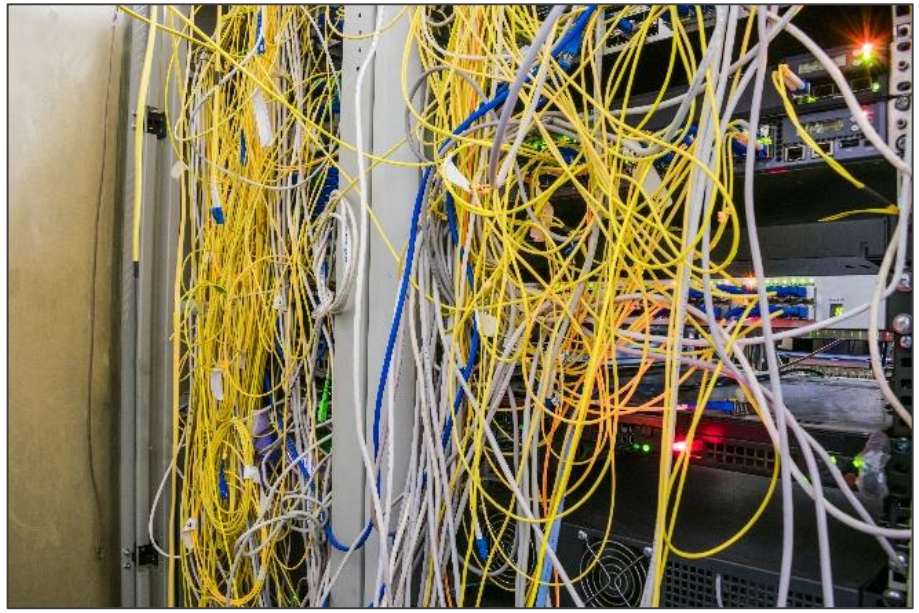
Maximiza el espacio, además de reducir el tiempo y el costo de la instalación.

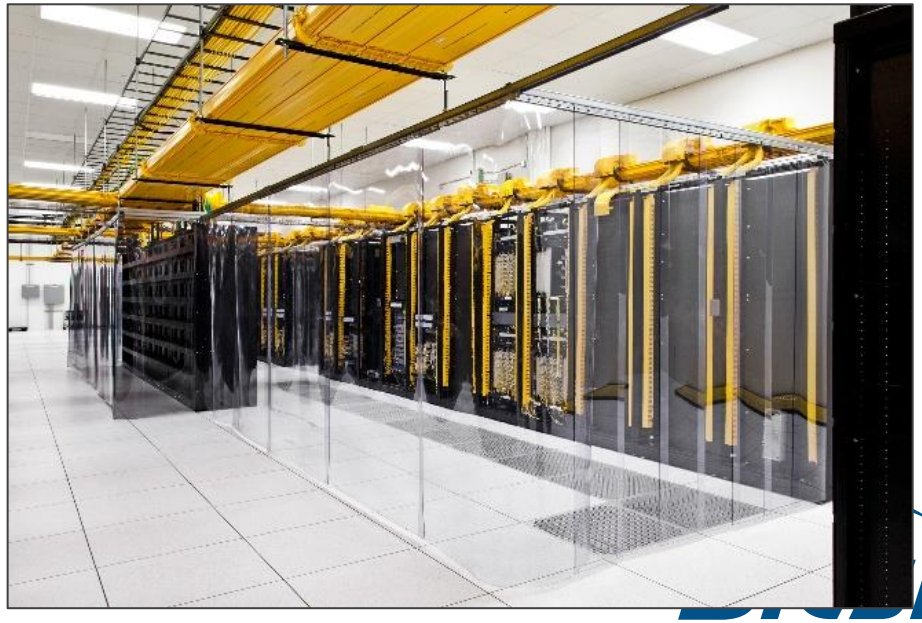
Proporciona el espacio adicional, necesario, para el crecimiento futuro.

Los movimientos, adiciones y cambios (MACs) pueden realizarse facilmente.

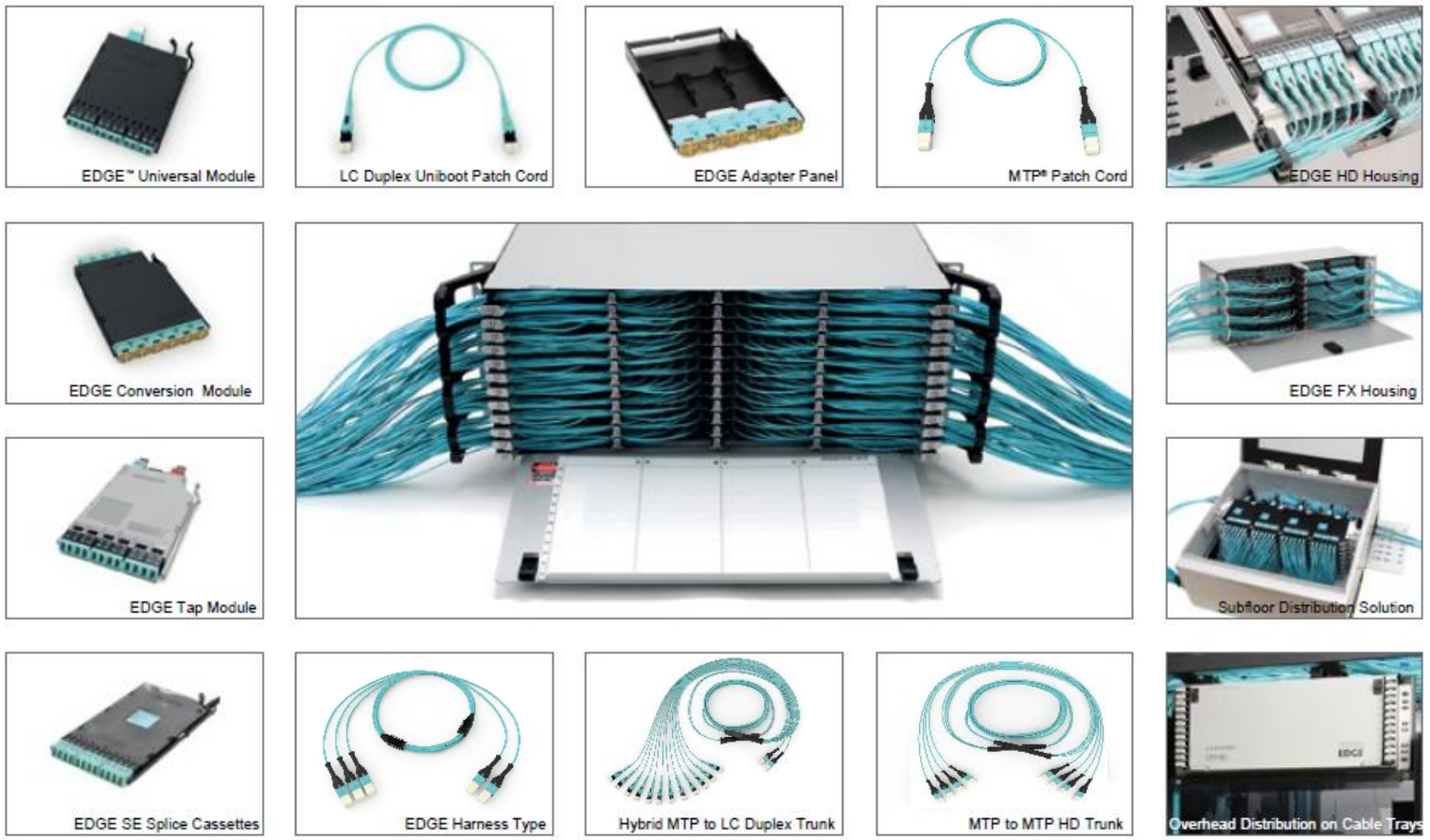
Soporta multiples generaciones tecnológicas

Una Infraestructura bien planificada puede durar entre 15 y 20 años





Descripción de la Infraestructura



La Infraestructura debe cumplir con los requisites de la evolución de los transceivers

Los sistemas Base (2, 8, 12, 24 F) tienen que tomar en cuenta la escalabilidad que requieren las tarjetas de switches con alto número de puertos

Las soluciones de conectividad deben coincidir con el transceiver y la densidad requerida

Los sistemas de infraestructuras flexibles deben ser modulares y adaptables

CORNING

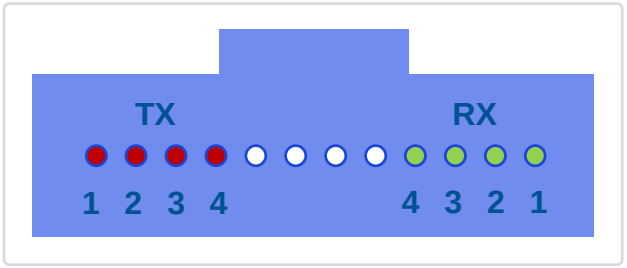
BICSI ENDORSED EVENT

Consideraciones de Diseño - Escalabilidad

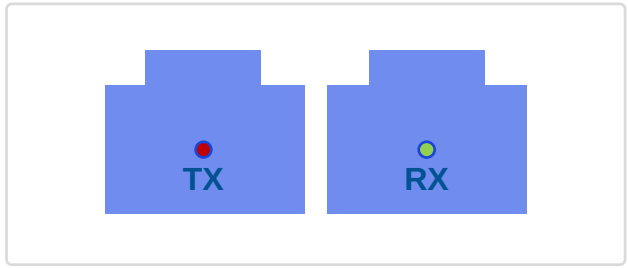
Escalabilidad de la red utilizando rompimiento de puertos

1x

40G
100G
400G



=



10G
25G
100G

4x

Los Puertos se pueden divider en subvelocidades individuales, lo que se conoce como rompimiento de puertos

El rompimiento de puertos permite escalar a velocidade superiors, utilizando el mismo equipo

Densidad

Costo

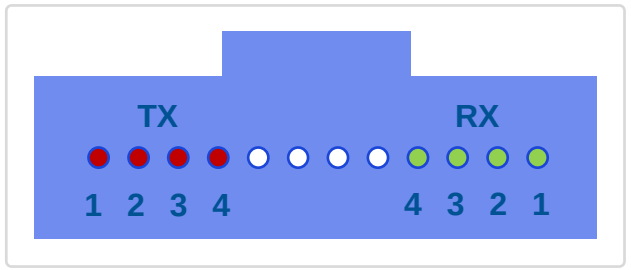
Energía

CORNING

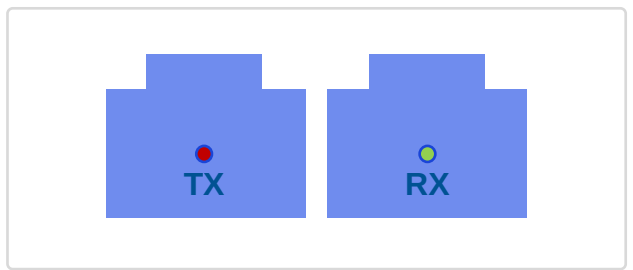
BICSI ENDORSED EVENT

Densidad Beneficios del Rompimiento de Puertos

1x 40G 100G 400Gb/s



=

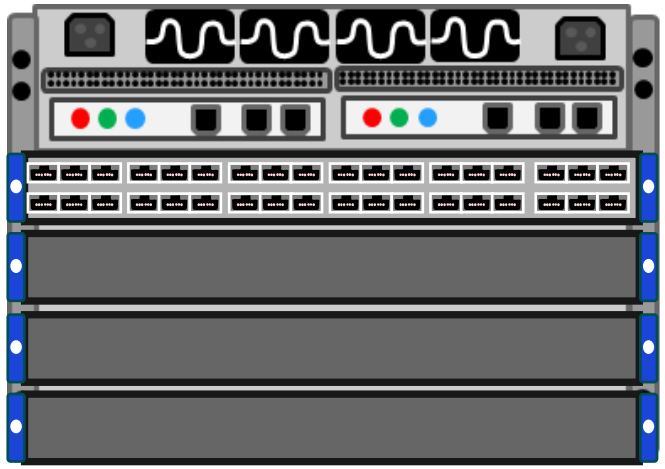


10G 25G 100G 4x

[Los Puertos se pueden divider en subvelocidades individuales](), lo que se conoce como rompimiento de puertos

1 Tarjeta de Red de 36 QSFP con puertos de 40G

1x

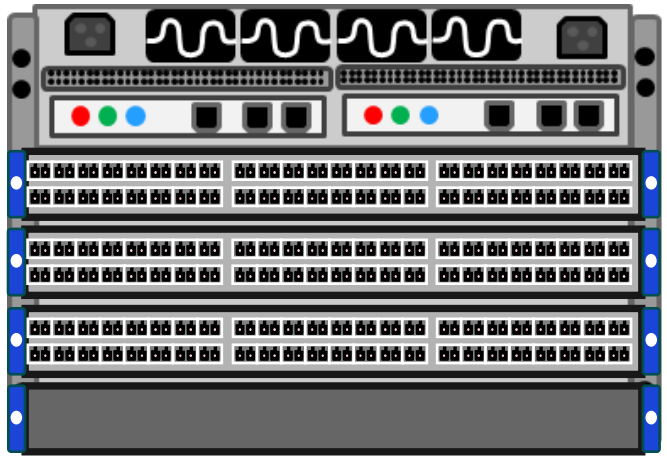


=

1 tarjeta de red x 36 puertos x **4 enlaces 10G** =
144 puertos de 10G

3 Tarjetas de Red de 48 SFP+ con puertos de 10G

3x



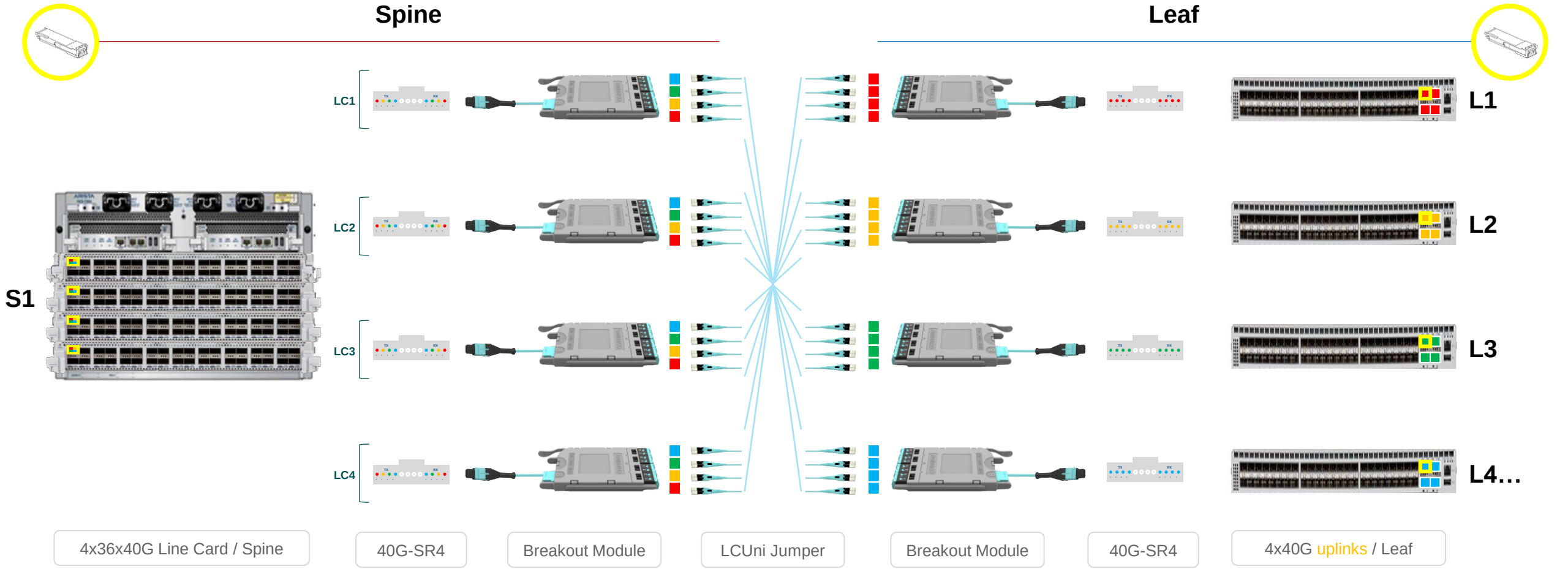
3 tarjetas de red x 48 puertos x **1 enlace 10G** =
144 puertos de 10 G

CORNING

BICSI ENDORSED EVENT

Data Center – Escalabilidad con Rompimiento de Puertos

Ejemplo: Ampliación de los Switches del Spine utilizando módulos de rompimiento, expandiéndose completamente a cuatro tarjetas de red por Spine

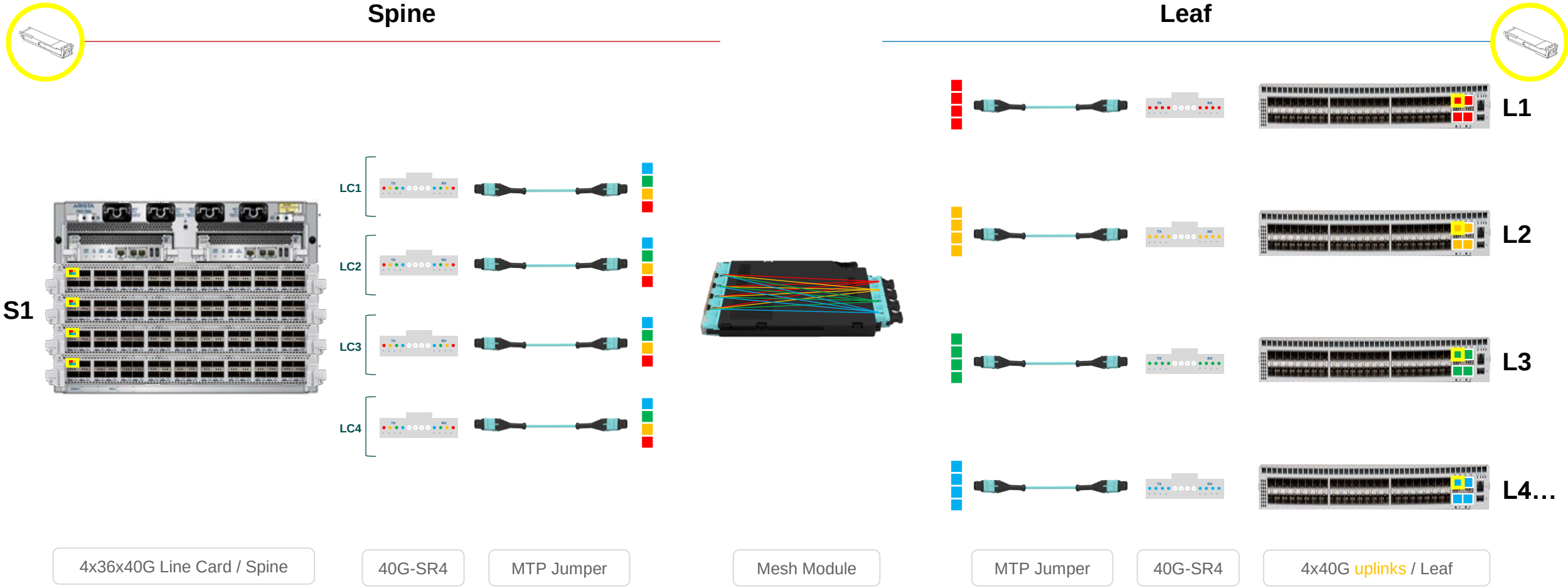


CORNING

Bicsi® ENDORSED EVENT

Data Center – Scalability with Port Breakout

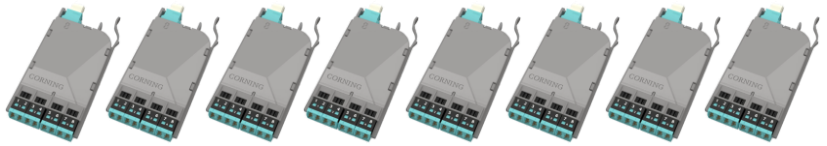
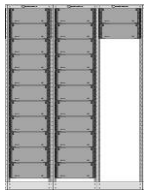
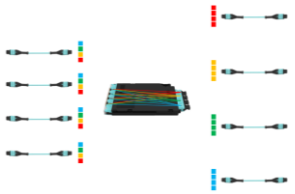
Ejemplo: Ampliación de los Switches del Spine utilizando módulo de mezcla, expandiéndose completamente a cuatro tarjetas de red por Spine



CORNING

Bicsi ENDORSED EVENT

Data Center – Módulos de Mezcla Prop-Valor

Característica	Beneficio	Valor	Rompimiento con Módulo	Rompimiento con LC
Ahorro en Costos	45% Menor costo de conectividad	Los Módulos de Mezcla tienen un menor costo de implementación que los módulos tradicionales LC Duplex		
Conectividad MTP	75% de Reducción de jumpers	La conectividad con MTP reduce el número de jumpers reqeuridos en el MDA vs los jumpers LC Duplex de rompimiento		
Densidad	75% de Reducción en espacio	La conectividad con MTP reduce significativamente la cantidad de espacio de rack requirido vs el rompimiento con LC Duplex		
Reduce la Atenuación del Enlace	10% menor IL	Menor pérdida por inserción, comparado con el rompimiento en LC Duplex	 1.35 dB	 1.5 dB

CORNING

Bicsi® ENDORSED EVENT

Beneficios del Monitoreo de la Red

Mitigación de Riesgo

Evite pérdidas causadas por fallas del sistema

Securidad

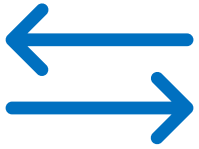
Detecta comportamientos sospechosos en su red

Tranquilidad

Mientras no reciba noticias de su sistema de monitoreo, sabrá que todo está funcionando

Respuesta Inmediata

Conozca los problemas rápidamente y antes de que se vuelvan críticos.



Resiliencia

Cambie a sus sistemas redundantes

Desempeño

Conozca los cuellos de botella en el desempeño, antes de que sus clientes se enteren

Actualizaciones

Los datos de desempeño le brindan la oportunidad de planificar e implementar actualizaciones antes de que la necesidad sea crítica

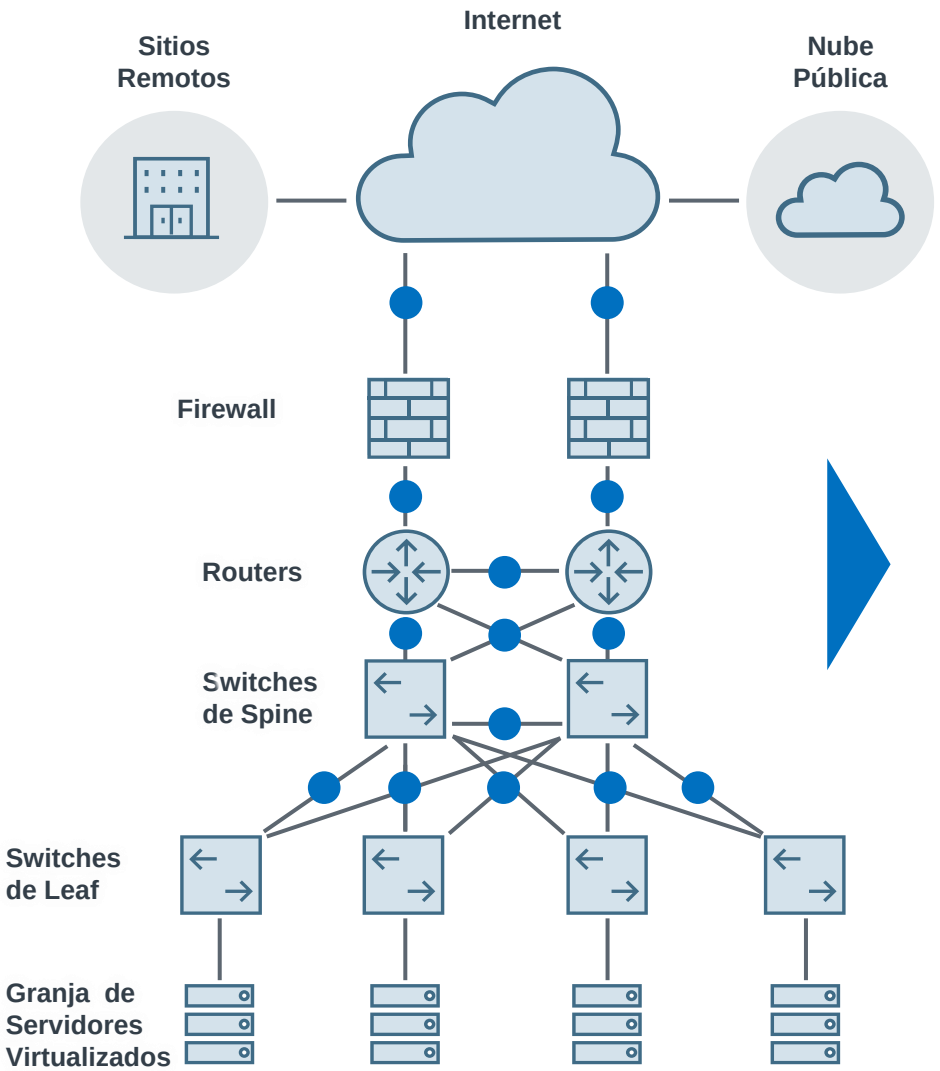
Proveedor

Compruebe si su proveedor cumple con el acuerdo de nivel de Servicio (service level agreement)

CORNING

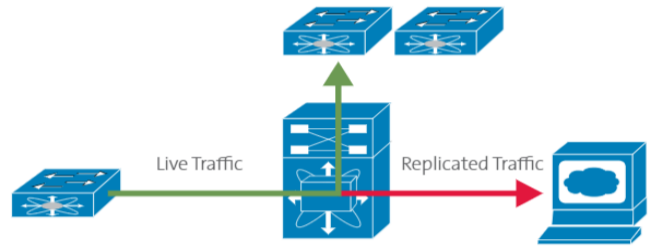
BICSI® ENDORSED EVENT

Descripción General del Monitoreo de Red

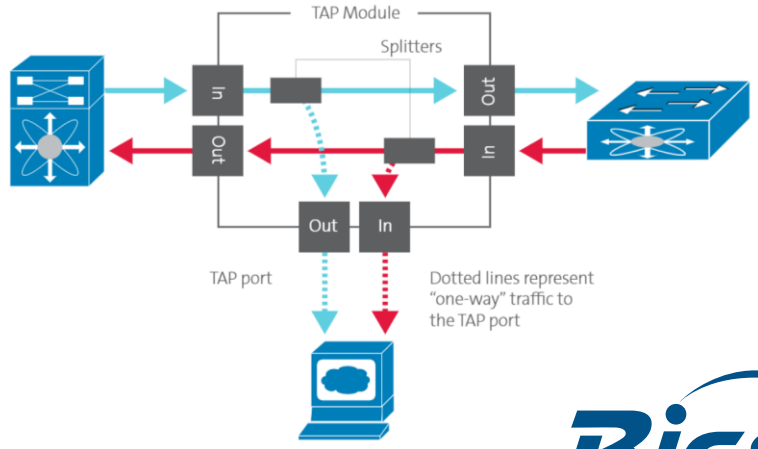


Monitoreo de Red

Tenología de SPAN



Tenología de TAP



CORNING

Bicsi® ENDORSED EVENT

Inteligencia Artificial y Machine Learning

CORNING

Bicsi
ENDORSED EVENT

¿Qué es IA/ML/DL?

Inteligencia Artificial

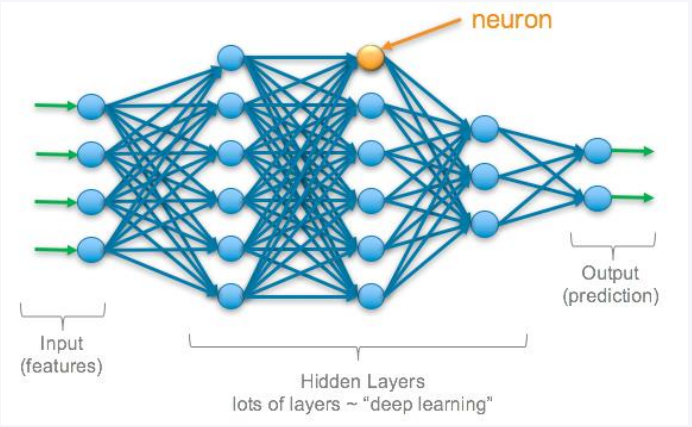
La teoría y el desarrollo de sistemas informáticos capaces de realizar tareas que normalmente requieren de inteligencia humana.

Machine Learning

Da a las computadoras la capacidad de aprender sin ser programadas explícitamente

Deep Learning

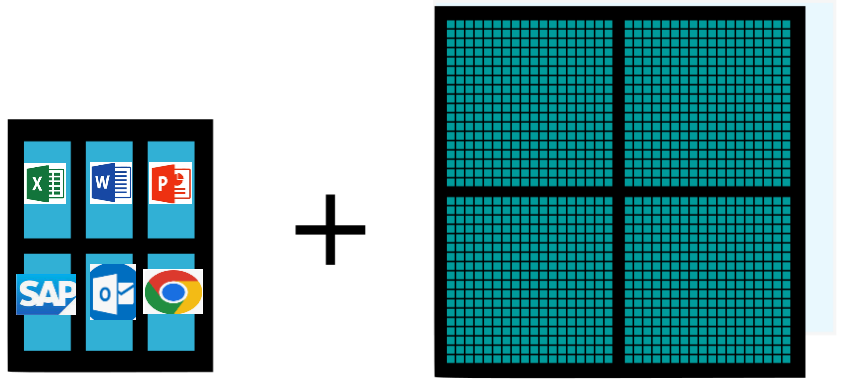
Algoritmos de aprendizaje automático, con estructura lógica, de estos algoritmos, similar a un cerebro, llamadas redes neuronales artificiales



CPU Versus GPU

- Una Unidad Central de Procesamiento (CPU) es el cerebro de una computadora, administra todas sus partes y ejecuta todo tipo de programas informáticos.
- Una unidad de procesamiento de gráficos (GPU) es un procesador especializado cuyo trabajo es acelerar la computadora para varias tareas específicas que requieren un alto grado de paralelismo, como generar (renderizar) imágenes u operaciones matriciales para el aprendizaje automático.

CPU	GPU
Paralelismo de Taréas	Paralelismo de Datos
Algunos núcleos “pesados”	Muchos núcleos “ligeros”
Tamaño de memoria alto	Alto rendimiento de memoria
Muchos conjuntos de instrucciones diversas	Algunos conjuntos de instrucciones altamente optimizados

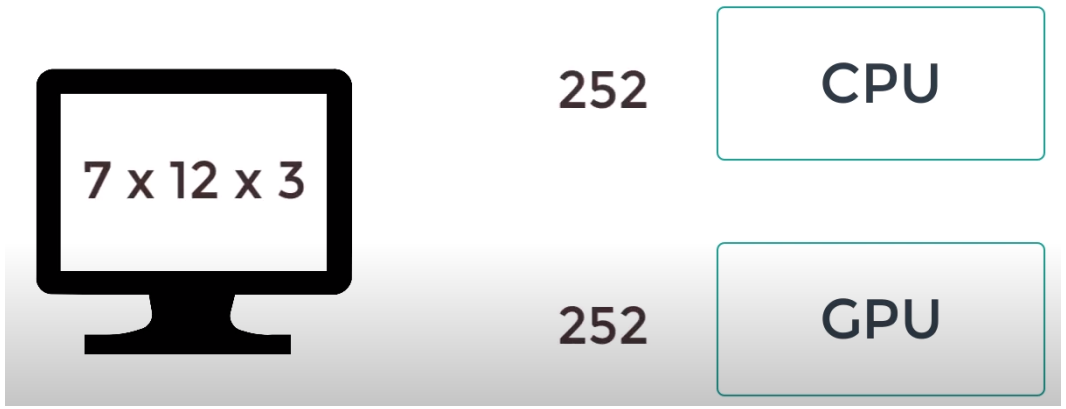


CPU
Multiple Cores

GPU
Thousands of Cores

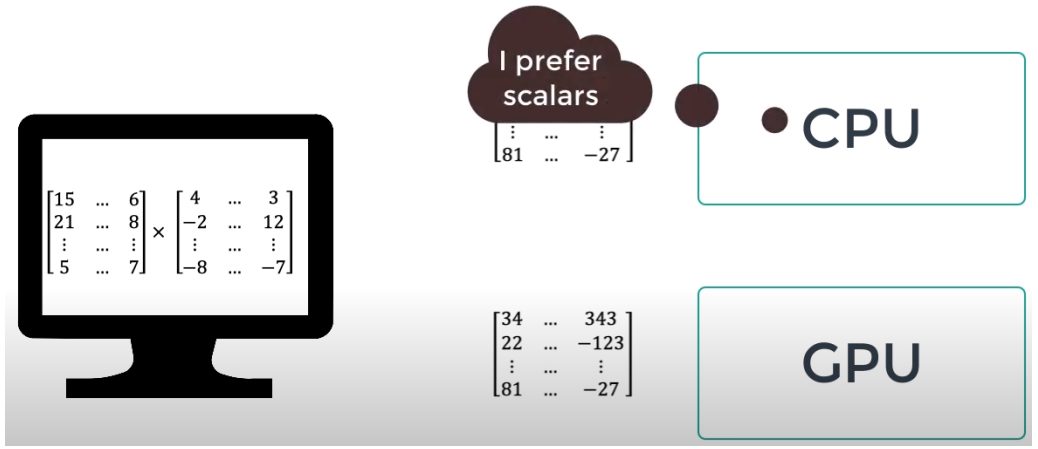
¿Cómo calcula una CPU comparado con una GPU?

Cálculo secuencial



Una CPU es más rápido que una GPU

Cálculo paralelo



Un GPU es más rápido que unaCPU

Fuente: https://www.youtube.com/watch?v=GRRMi7UfZHg

¿Qué está impulsando la necesidad del Cómputo de Alto Desempeño (HPC)?

Inteligencia Artificial (IA)

- Incluye un conjunto de tecnologías que permiten a las máquinas y/o computadoras imitar la inteligencia humana.
- Las tecnologías que caen bajo el paraguas de la IA incluyen el procesamiento del lenguaje natural y el aprendizaje profundo (modelos de lenguaje amplio/LLM) que comprenden texto humano.
- Los modelos de IA son programas que analizan datos y hacen predicciones. LLM puede tener más de mil millones de parámetros (variables).
- Un número cada vez mayor de industrias están adoptando tecnologías de inteligencia artificial para desarrollar nuevas formas de hacer negocios y aumentar la eficiencia.

Machine Learning

- Algoritmos que utilizan las máquinas para analizar grandes cantidades de datos
- Un modelo de aprendizaje donde se analizan los datos, se identifican patrones y se hacen predicciones.
- Los servidores con GPU (Unidades de procesamiento gráfico) están mejor equipados para realizar el procesamiento necesario para entrenar los modelos de IA y mejorar los resultados del aprendizaje automático.
- Las GPU están especializadas y pueden realizar miles de operaciones a la vez (procesamiento paralelo)
- Se puede lograr un aumento del rendimiento y la capacidad de procesamiento conectando las GPU en un clúster.

Bard

Generative AI

Virtual Assistant

Social Media

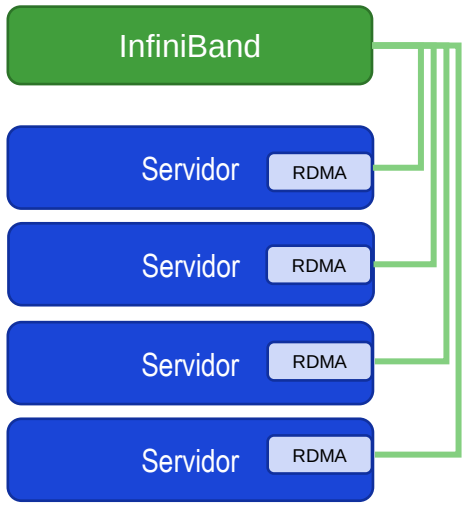
tv

prime video

Streaming Content

Diseño de red para IA/Back End

Diferencia entre los servidores usados para el Front End v. Back End

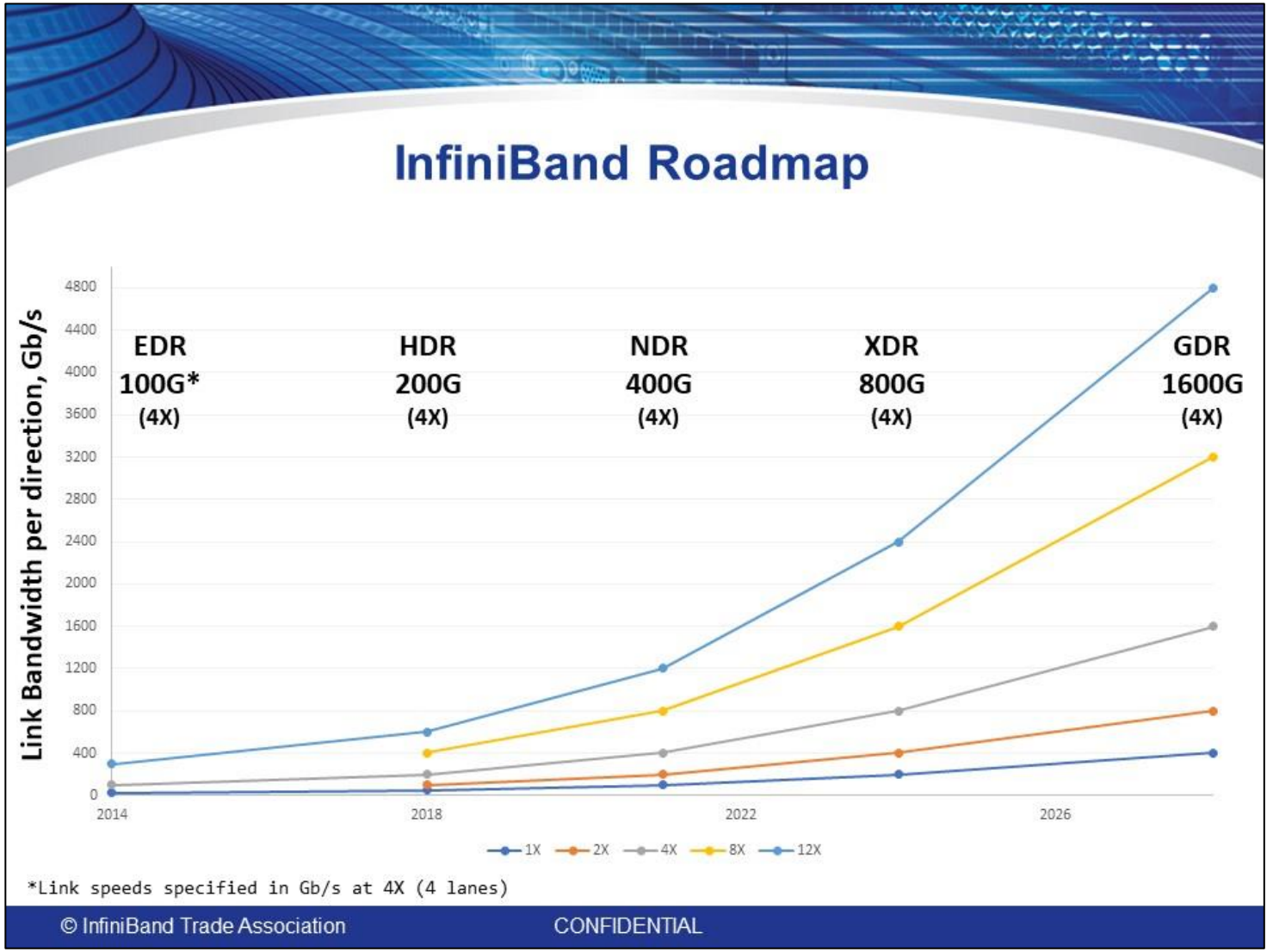


Servidores de Alto Desempeño & Servidores de IA

Consideraciones de Diseño

- Latencia ultra-baja: redes 1:1 para permitir que los algoritmos de IA realicen los cálculos necesarios
- Cada servidor de alto desempeño tendrá varias GPU's. Cada GPU deberá estar interconectada para cumplir los objetivos de la red de IA.
- Redes InfiniBand versus Ethernet
- Velocidades de los puertos de 400 Gbs, 800 Gbs, 1,6 Tbs. Interfaces de transceiver MTP y pulido APC

Roadmap de Velocidades de InfiniBand



CORNING

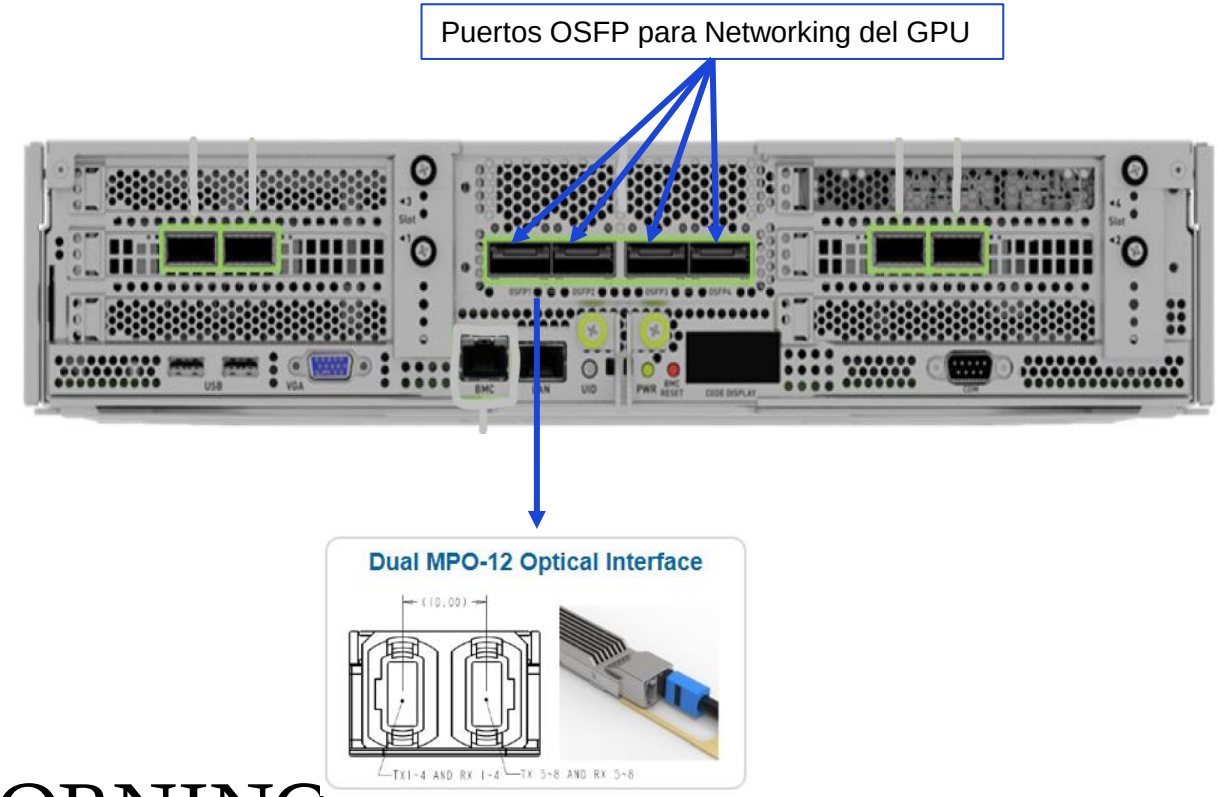
Bicsi
ENDORSED EVENT

Conexiones GPU/CPU de un Servidor de IA

Un servidor de IA de alto rendimiento está equipado con múltiples GPU's interconectadas para realizar el procesamiento de modelos de IA.

Los transceivers ópticos de los GPU's tienen un formato OSFP con conexiones duales MTP-8 (800 Gbs o 2 x 400 Gbs)

Se utilizan puertos QSFP separados para proporcionar redes para almacenamiento o administración OOB



Conexiones Totales en MTP para Servidor de IA

- 8 GPU's
- 4 Puertos OSFP
- 8 MTP's de 12 (8 fibra/conector)
- GPU Networking: 64 fibras por servidor

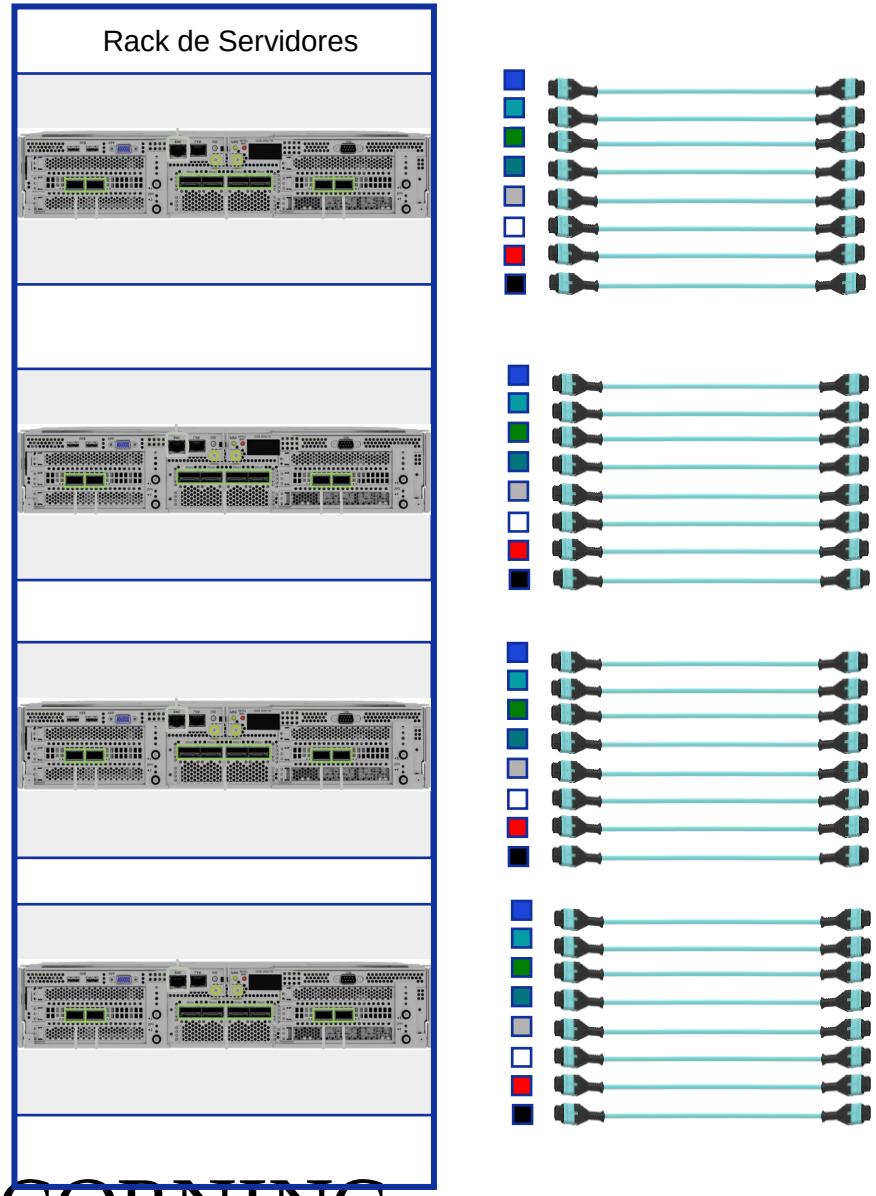
- Un servidor basado en CPU solo puede tener 1 o 2 puertos SFP/QSFP

- La conectividad de fibra para un Servidor de IA/Alto Rendimiento puede ser **5 x** mayor que la de un servidor de centro de datos estándar

CORNING

Bicsi ENDORSED EVENT

Construyendo el Cluster de IA



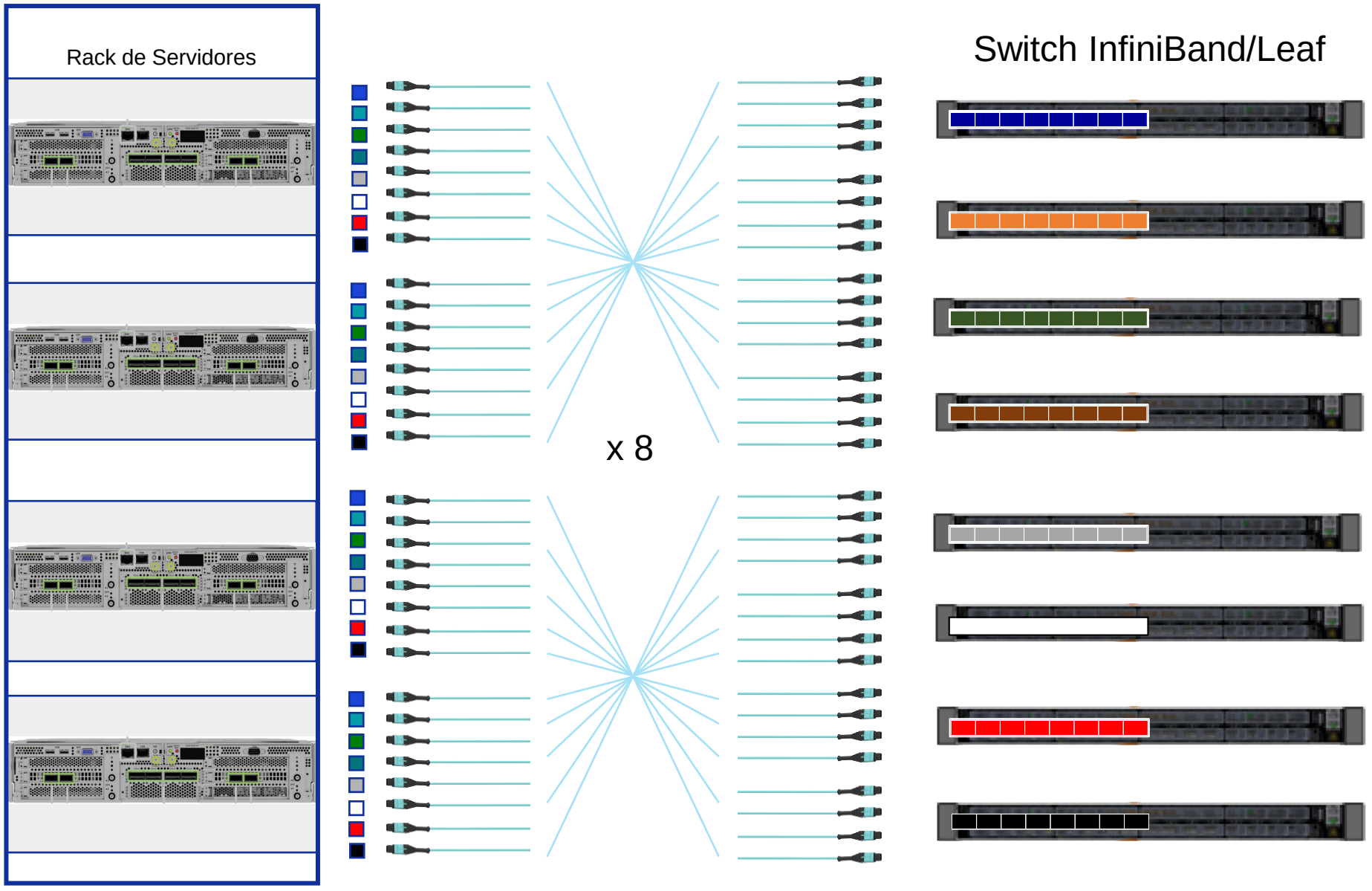
Switch InfiniBand Quantum-2

- Se pueden instalar hasta 4 servidores de Alto Rendimiento/IA en un rack de red
- Los requisitos de energía y refrigeración limitarán la cantidad de servidores de IA en un rack (~40Kw/rack)
- Los clusters pequeños pueden usar jumpers individuales para conectar el switch MOR/EOR

CORNING

Bicsi® ENDORSED EVENT

Construyendo el Cluster de IA – Fila Completa del Servidor



- Swiches NVIDIA InfiniBand soportan 32 puertos OSFP
- Cada puerto OSFP soportará 800Gb/s
- La base del switch Quantum-2 es de 64 puertos x 400Gbps
- Con una red 1:1 completa y sin bloqueo, cada switch admitirá 32 puertos de downlink y 32 puertos de uplink
- Radix habilita una fila de servidores para permitir 32 Servidores de IA

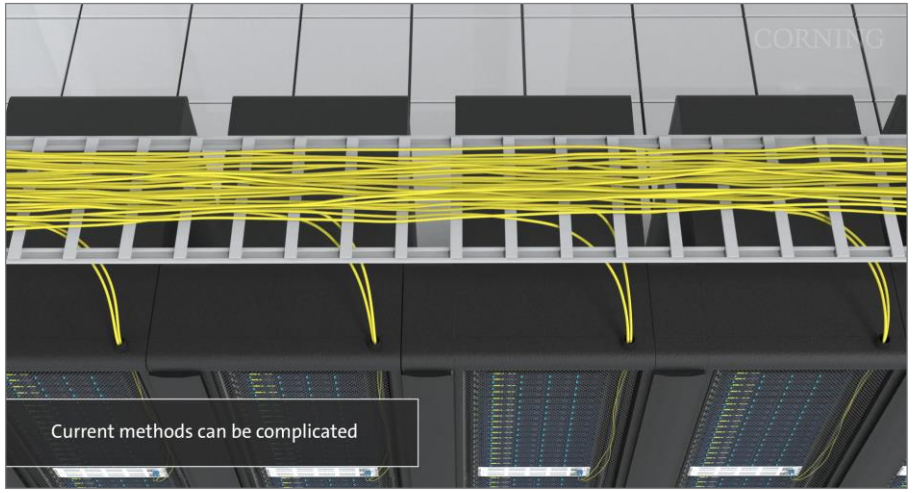
CORNING

Bicsi® ENDORSED EVENT

Consideraciones de cableado al construir un Cluster de IA

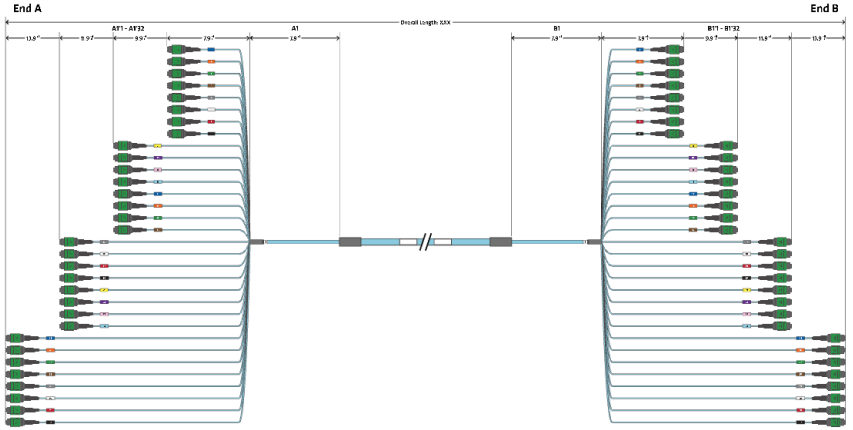
Jumpers Individuales MTP (jumpers de 8 Fibras)

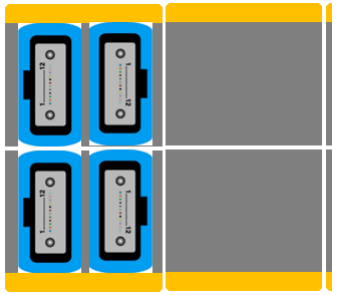
- Menos Costoso
- Más Complejo
- Se necesita adaptar a multiples longitudes
- Se requiere mayor mano de obra para juntar con Velcro los jumpers antes de la instalación dentro de la charola de cables
- Congestionamiento en el rack: OD de un Jumper MTP es de 3.6mm



Bundled MTP/8 Fiber Jumpers

- Más Costoso
- Más Fácil de instalar
- Los conectores MTP Pro son más fáciles de conectar/desconectar
- Solución diseñada con cables de rupture escalonados
- Mitiga la congestión de cables dentro del rack de IA : OD de los cables MTP es de 2.0mm

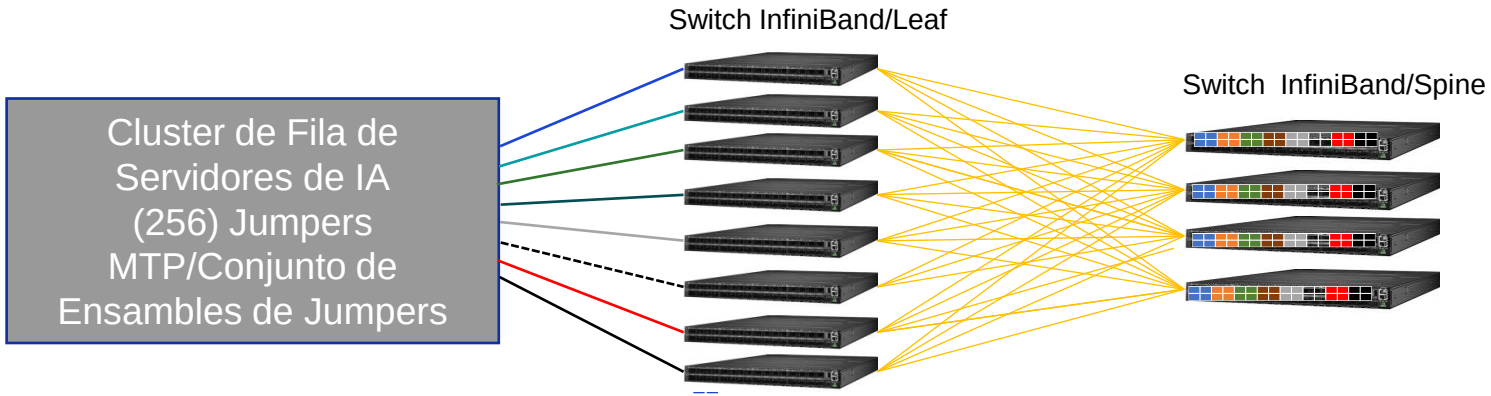




CORNING

Bicsi® ENDORSED EVENT

Networking de Data Center con SuperPods para IA



Para interconectar las 256 GPU's, en un único clúster de IA, se necesitarán conexiones de 400 Gbps a través de interfaces MTP12.

Conexiones físicas para el Switch Spine:

- 256 conexiones MTP12 desde InfiniBand/Leaf
- 4 switches InfiniBand porcionaran conectividad
    - 32 puertos OSFP
    - 64 conexiones NDR (2 x MTP12)

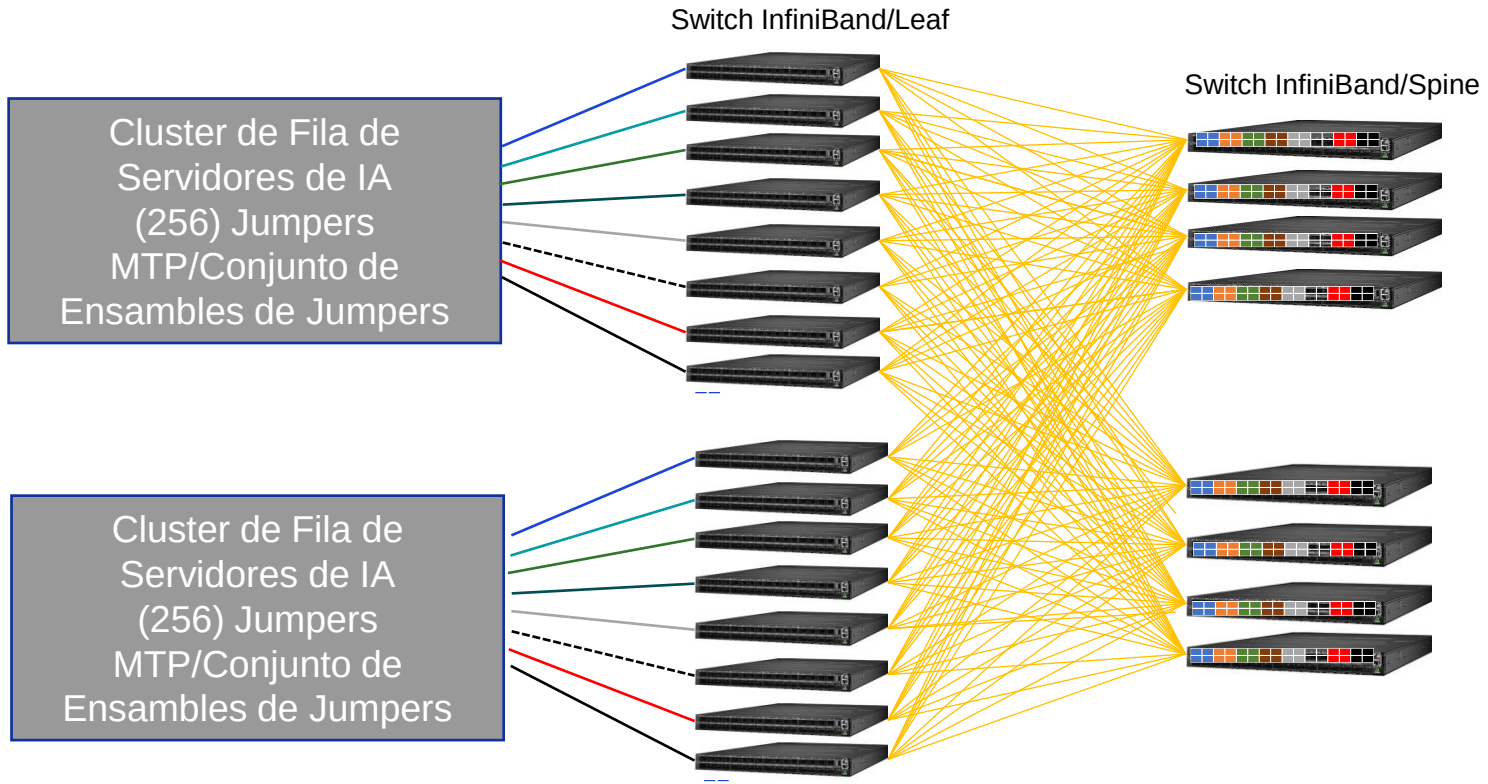
Tipo de Fibra

- La Fibra Multi-Modo, y sus componentes, será el diseño más rentable, desde el servidor hasta el switch
- Fibra Mono-modo de switch a switch

CORNING

BICSI ENDORSED EVENT

Networking de Data Center con SuperPods para IA



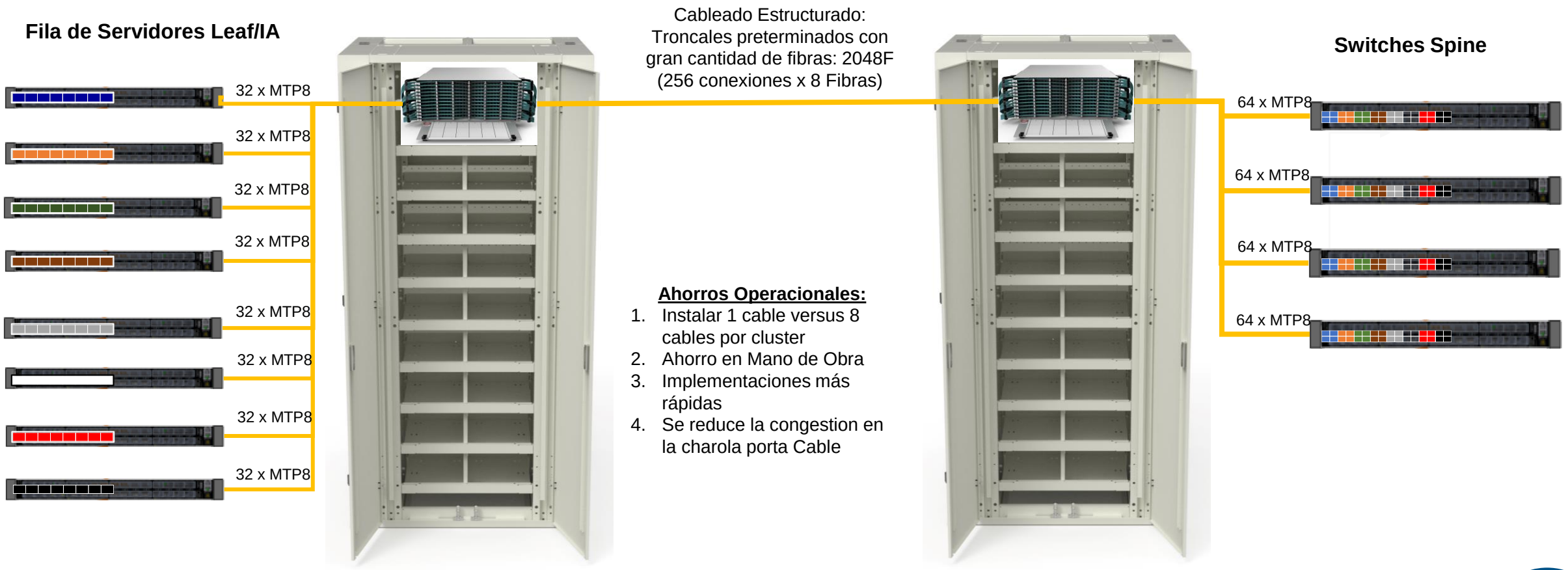
<u>Cada nuevo Cluster agregará</u>:

- 4 Nuevos Switches de Spine
  - 4 nuevos switches InfiniBand proporcionan conectividad
    - 32 puertos OSFP
    - 64 conexiones NDR (MTP12)
    - 256 conexiones totales

- Agregar un nuevo clúster también requeriría cambios :
  - Cableado en el Switch de Spine
  - Necesidad de mover conexiones a un nuevo conjunto de Switches de Spine

- Los desafíos operativos aumentan a medida que aumenta el número de filas de servidores para IA
  - Tiempo de Ingeniería
  - Tiempo de mano de obra
  - Documentación
  - Pruebas

Cableado Estructurado para topología Spine/Leaf para Clusters de IA



Ahorros Operacionales:

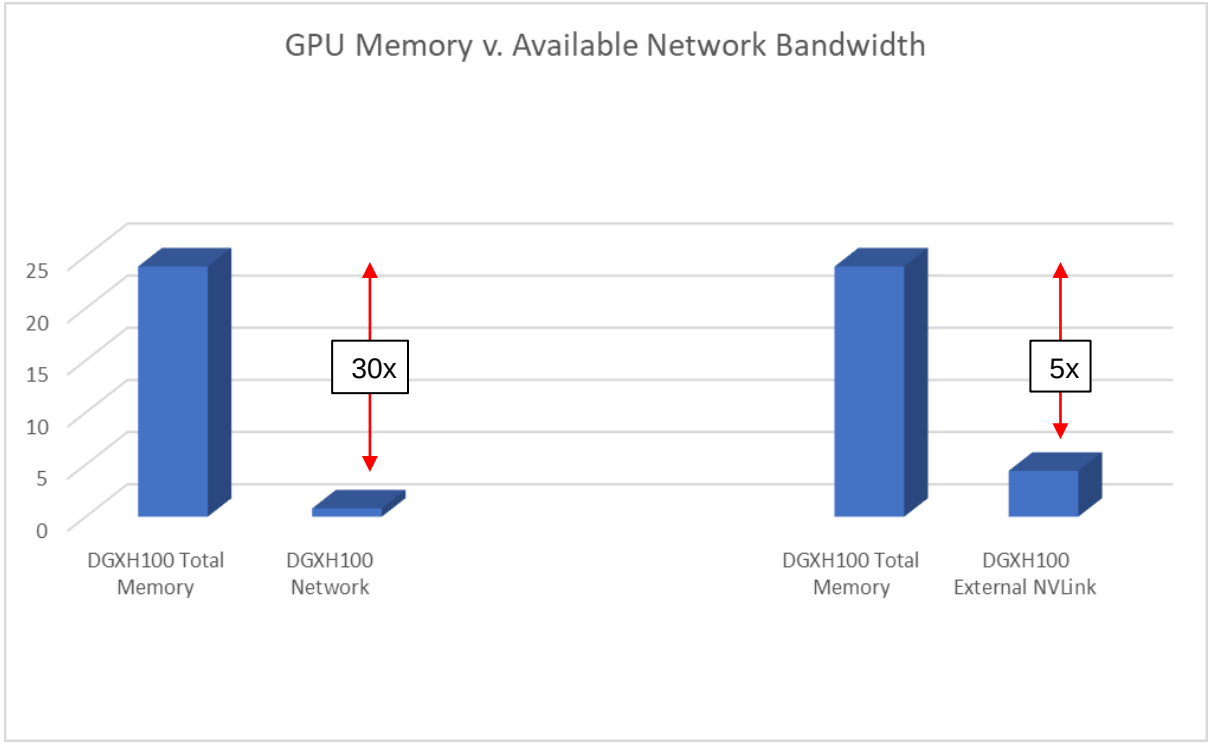
1. Instalar 1 cable versus 8 cables por cluster
2. Ahorro en Mano de Obra
3. Implementaciones más rápidas
4. Se reduce la congestion en la charola porta Cable

CORNING

Bicsi ENDORSED EVENT

Necesidad de una 2ª Red de Back End

- Los Servidores actuales de NVIDIA para IA tienen un ancho de banda de memoria total de 24 Terabytes/seg
- Las interfaces de GPU internas proporcionan un ancho de banda de red de 800 Gb/s... creando una brecha de memoria (gap 30x)
- Conectar la GPU externamente sobre una red de fibra puede reducir esta brecha a 5 veces
- La expansión de los modelos de IA requiere reducir la brecha para garantizar una baja latencia



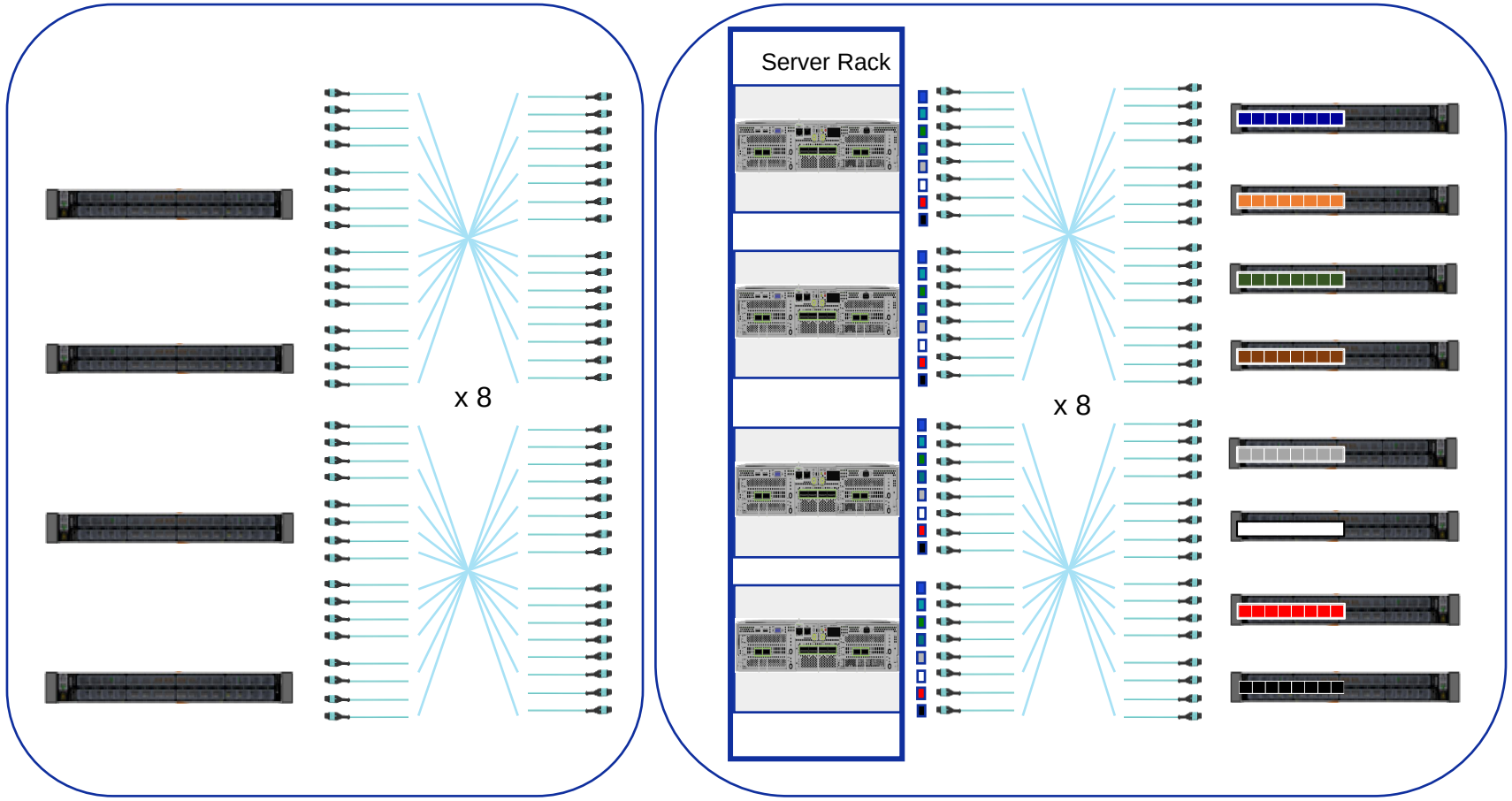
2ª Red de Back End: 100% aumenta la conectividad de Fibra

2a Red de Back End

- La próxima generación de servidores NVIDIA tendrá más puertos de I/O para conectividad

- (8) GPUs por Servidor
- (32) Servidores por Cluster
- (256) jumpers MTP8 adicionales se requieren para crear una red NVLink externa

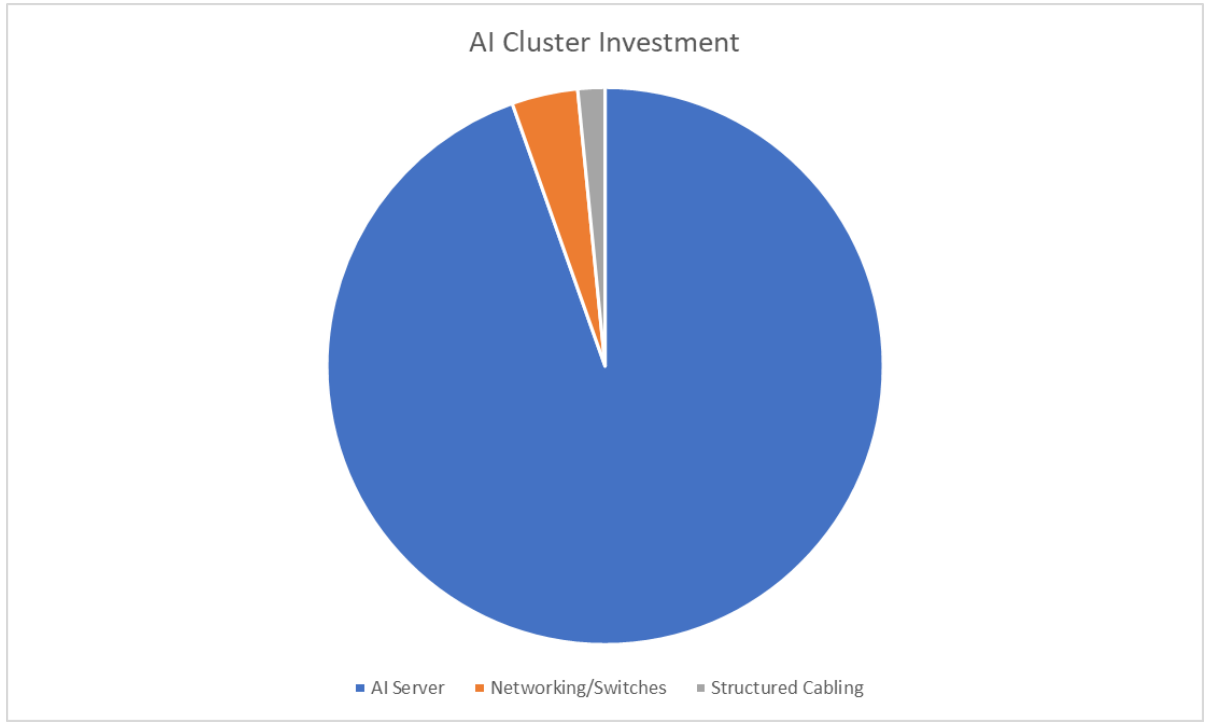
- Ilustra además la necesidad de considerar conjunto de ensambles de jumpers o cableado estructurado



CORNING

Bicsi ENDORSED EVENT

Desglose de Costos



- Más del 95% del costo de la red, para el clúster de Servidores de IA, está relacionado con el equipo activo
- El gasto en Energía y Refrigeración también serán importantes
- La conectividad de fibra facilita la velocidad de la red y la escalabilidad del procesamiento.
- La barra es baja para justificar el gasto de un cableado estructurado premium

Resumen

- Los requisitos de fibra para redes de Computo de Alto Desempeño (HPC) e IA, son 5 veces mayores que los de la red de producción de centros de datos tradicionales
- El conjunto de ensambles de jumpers Punto – Punto pueden acomodar clusters más pequeños; los escalables más grandes necesitan soluciones de cableado estructurado
- Roadmaps para transceivers Ethernet e InfiniBand se ampliarán con backbones de fibra Base8
- Los elementos del sistema de cableado estructurado (TAPs pasivos, Rompimiento de Puerto) permitirá a los operadores del data center obtener más valor de la infraestructura de fibra.