

# The Smart Route to AI Workload Efficiency

## The Challenge - Efficient AI Workload Deployment

The unique properties of AI workloads present a formidable set of challenges when deploying AI infrastructure. Ensuring optimal performance in high-speed data center environments can be met with various challenges.

- **Path Selection:** Critical for reducing Job Completion Time (JCT), effective path selection ensures that resources do not remain idle while waiting for the final packet to be sent and received.
- **Resource Utilisation:** Current solutions often involve purchasing additional GPUs to boost performance rather than optimising existing resources, leading to inefficiencies.
- **Network Congestion:** Congestion can severely hamper performance, making it crucial to have intelligent routing that can adapt in real-time.

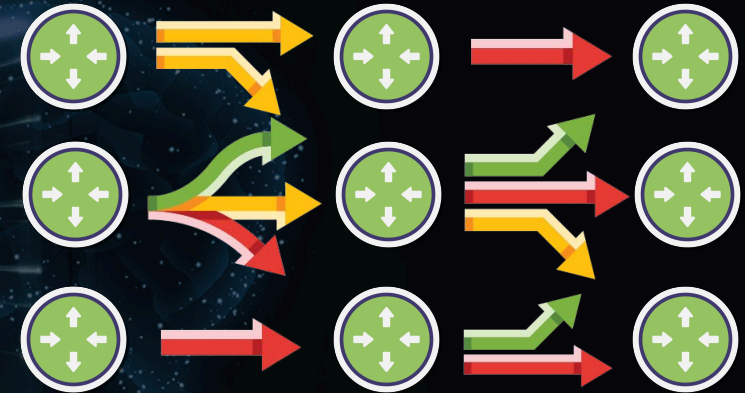
## AI Infrastructure

To address these challenges, modern AI networks utilize different high performance network technologies:

**Telemetry Assisted Ethernet:** Uses real-time data and analytics to dynamically adjust routing paths, avoiding congestion and ensuring efficient data flow. By optimizing paths in real time, it reduces latency and enhances overall network performance, critical for demanding AI workloads.

**Fully Scheduled Fabric:** Ensures deterministic networking with guaranteed bandwidth and predictable performance. By transmitting data packets at scheduled times, it reduces collisions and provides consistent low-latency communication, ensuring essential AI processes receive the necessary bandwidth and prioritization for optimal efficiency and reliability.

## Job Completion Time - Only as fast as your slowest route



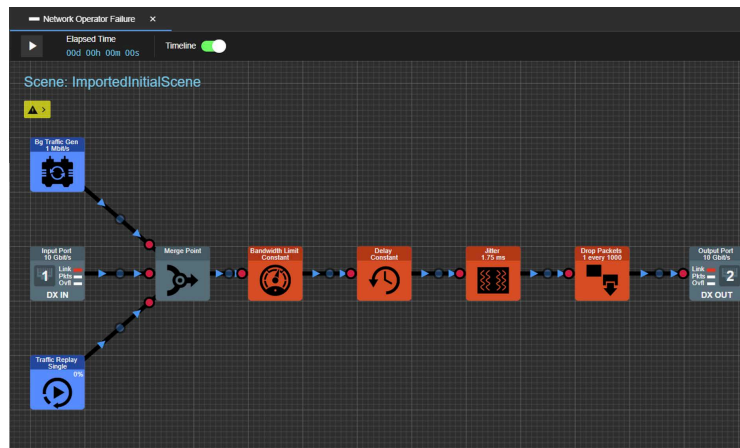
## Calnex SNE-X, the 400GbE AI network performance twin test solution

The Calnex 400Gbps network emulator, SNE-X, is the first of its kind, providing a comprehensive solution for device and cluster testing, as well as AI network fabric twinning.

### Impairments can be injected into packet streams to cause adverse effects.

Emulating network impairments in a controlled and repeatable environment helps identify and mitigate issues in AI infrastructure, ensuring optimal performance and reliability.

- **Bandwidth Throttle** can be used to create congested links.
- **Delay** will add latency to increase the JCT.
- The **Packet Modifier** can corrupt or edit contents of a packet, for example, the ECN field.
- **Reorder** will have packets arrive out of sequence.



### Calnex SNE-X Highlights:

- 400GbE, 100GbE, 50GbE, 25GbE, 10GbE and 1GbE interfaces supported
- Web-based user interface
- RESTful API for easy remote control in your automation environment
- Multi-user
- Connect any-port to any-port without limitations

Mimic ever-changing real-world network conditions in a controlled, repeatable manner. This allows for comprehensive testing and optimization, ensuring that workload performance is accurately evaluated under various scenarios.