2023 embedded VISION SUMMIT

Using a Neural Processor for Always-Sensing Cameras

Sharad Chole

Chief Technologist, Co-founder

Expedera

expedera

Always-sensing Cameras: What and Why

- Like always-listening Siri, always-sensing enables a more natural and seamless user experience
- Quality and richness of camera data require much more processing than audio implementations
  - Like always-listening, always-sensing must be processed locally
- Technical requirements:
  - ~500 GOPS to 1 TOPS NPU
  - Ultra-low power
  - Ultra-small area
  - Multiple models

About Expedera

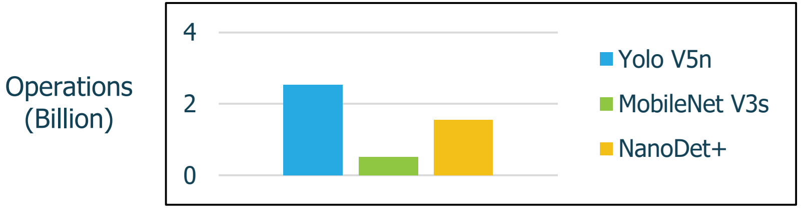
- Optimized edge AI inference IP solutions based on revolutionary packet architecture, application-configured for our customers
- Silicon Valley startup founded in 2018
  - 3 R&D centers, numerous patents
- Broad, worldwide deployments
  - 10M+ devices in-field
  - Over 200 ExaOps (200,000,000,000,000,000,000 operations/second) deployed by our customers
  - Soft IP: designs in multiple leading-edge nodes

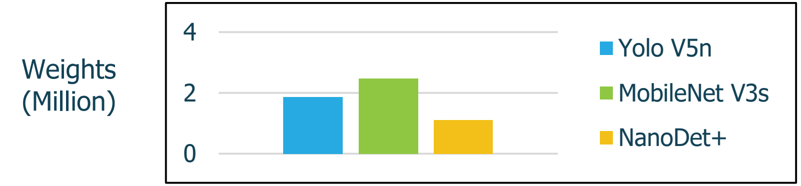
Always-sensing Cameras: What and Why

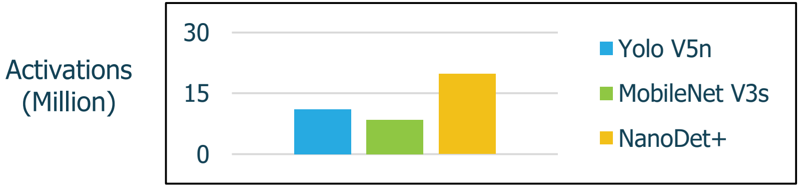
- Like always-listening Siri, always-sensing enables a more natural and seamless user experience
- Quality and richness of camera data require much more processing than audio implementations
  - Like always-listening, always-sensing must be processed locally
- Technical requirements:
  - ~500 GOPS to 1 TOPS NPU
  - Ultra-low power
  - Ultra-small area
  - Multiple models

NPUs and Always-Sensing Are an Ideal Match

- NPUs are used in wake-word (audio) applications - why can't the same approach apply to video?
- Workloads – Edge-friendly vision NNs
  - ISPs/DSPs aren't designed for this sort of specialized processing
- We asked ourselves – rather, customers asked us – can we apply the same low power, small, targeted workload approach to video?
  - We have; otherwise, I wouldn't be giving this talk



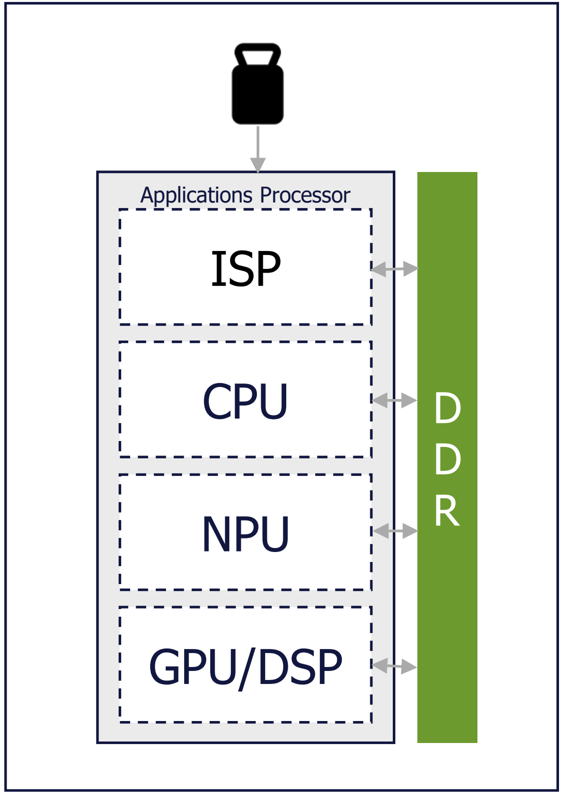




Always-sensing Use Cases

- Secure access: facial recognition
- Power management: “find a face” detection to turn on/down/off display
- Gesture recognition: Innovative UX control and operability
- Motion detection: bandwidth-friendly security
- Object detection: capturing events or triggers
- Privacy: “shoulder surfing” alerts

Existing Solutions are Limited



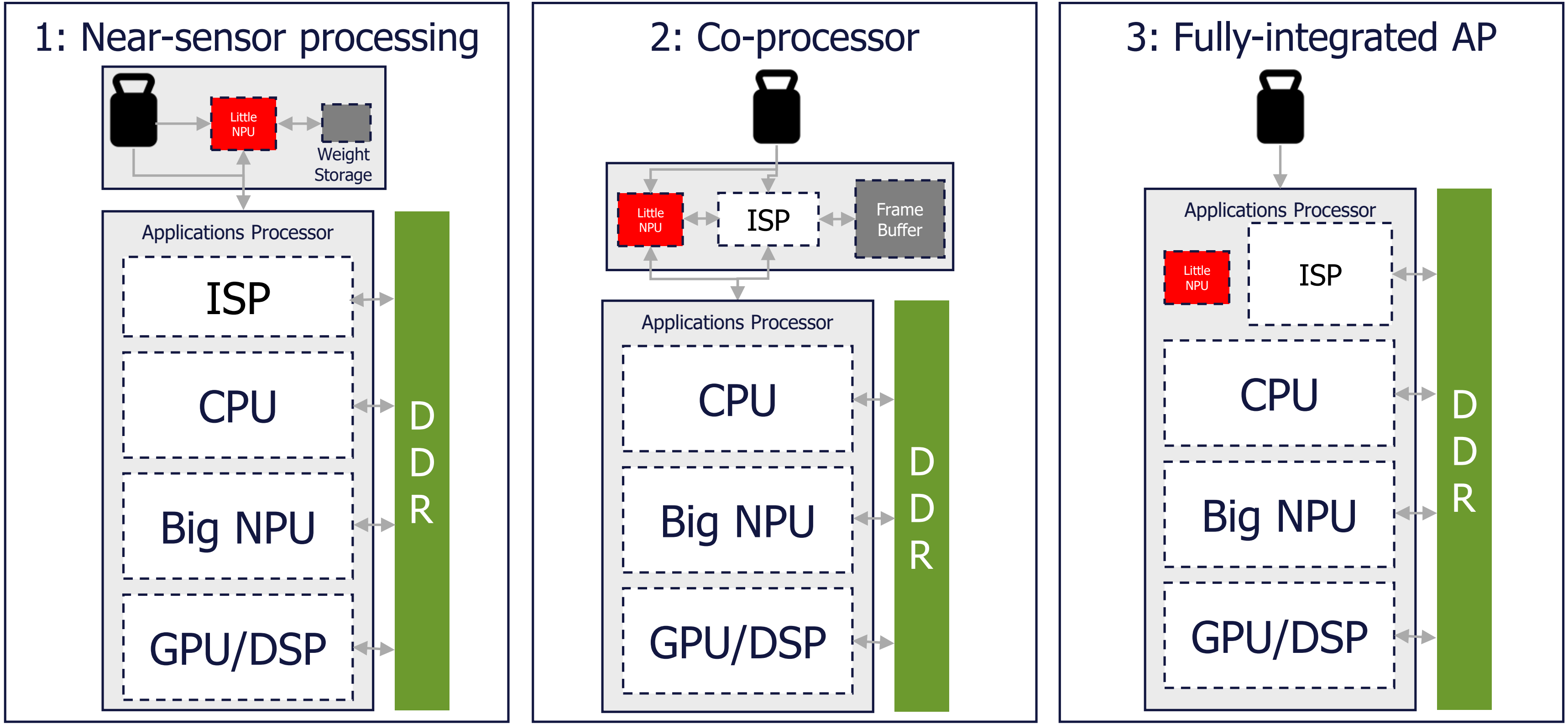
- NPUs, GPU/DSPs, and ISPs are standard in Application Processors
- The system NPU – the "Big" NPU – is not a good match for always-sensing
  - Excessive power consumption
  - Privacy, memory & data security concerns
  - Contention by multiple applications
  - "Hitting a nail with a piledriver"

The Case for a Little NPU: Ideal Always-sensing

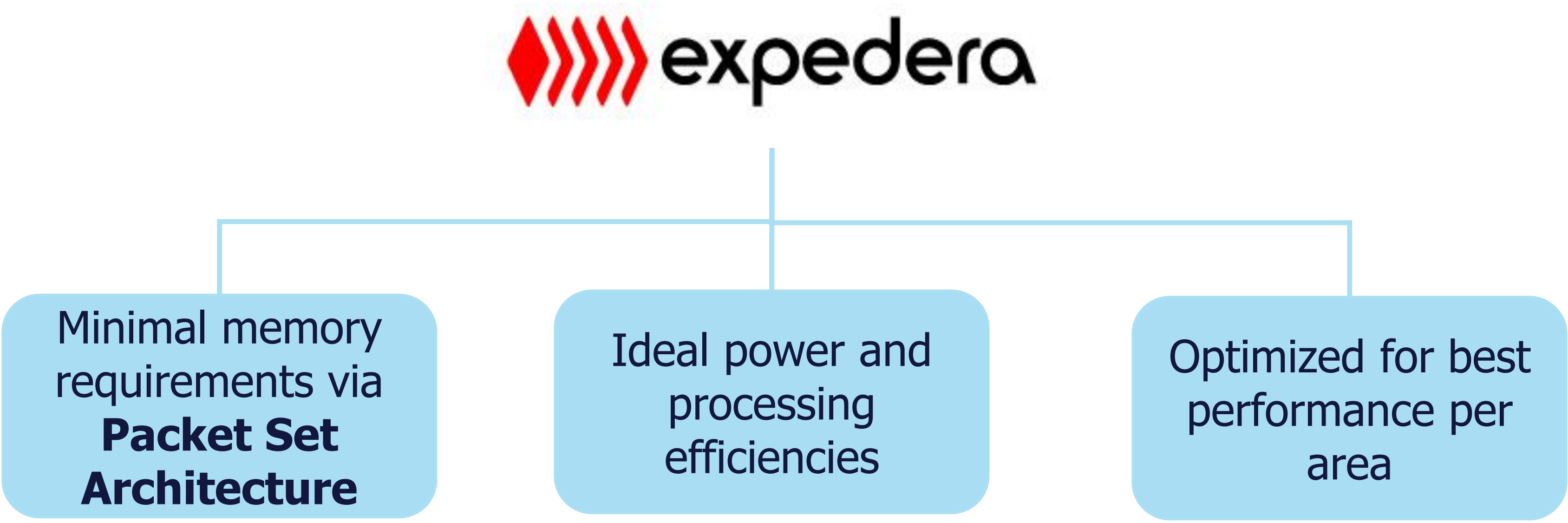
- Purpose-designed for smallest area, lowest power consumption, and target networks
  - Area: reduces cost
  - Latency: programmable, guaranteed FPS
  - Power consumption: conserves battery life
- No requirement for external memory
  - Keep processing and data storage local
  - Increased power efficiency
  - Better security and privacy

Specification	Big NPU	Little NPU
Area	~5-8X	1X
Subsystem Size	>10X	1X
Latency	Contention w/ activity	Deterministic
Data Exposure	System DDR	Within always-sensing subsystem

Big/Little NPU Architecture Options



Expedera’s NPU Architecture



expedera

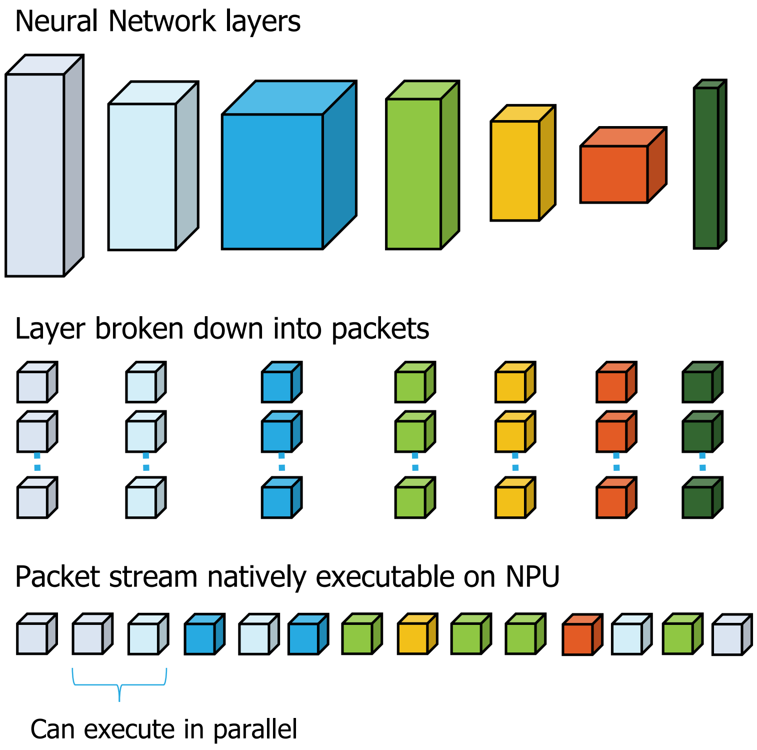
Minimal memory requirements via **Packet Set Architecture**

Ideal power and processing efficiencies

Optimized for best performance per area

Packets: A Radical Approach to AI Inference

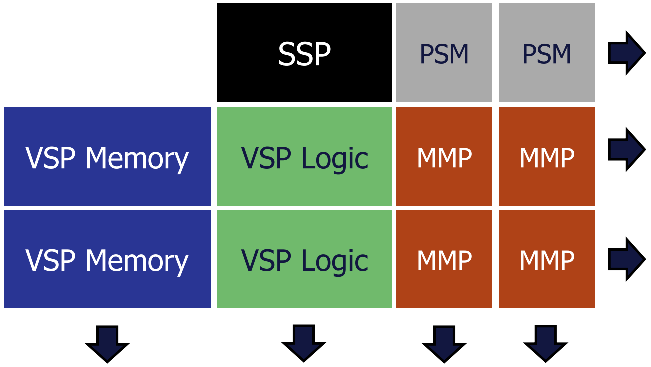
- Packet - aggregate of work with a notion of dependencies and deterministic execution
  - A contiguous fragment of a NN layer with the entire context of execution: layer type, attributes, priority
- Packets manage activations better/more intelligently
- Results: minimum number of moves without hurting accuracy
  - Greatly increases performance while lowering power and area requirements


Neural Network layers
Layer broken down into packets
Packet stream natively executable on NPU
Can execute in parallel

© 2023 Expedera Inc.

Expedera Origin^{TM} NPU Engine

- Revolutionary packet-based architecture reduces design and implementation complexity while improving real-world performance
- Hardware and software are designed together
  - Solves complex software in hardware – compiler co-designed and optimized for unique architecture
- Just-in-time memory management implements a unified compute pipeline
- Expedera-optimized for customer use case(s)



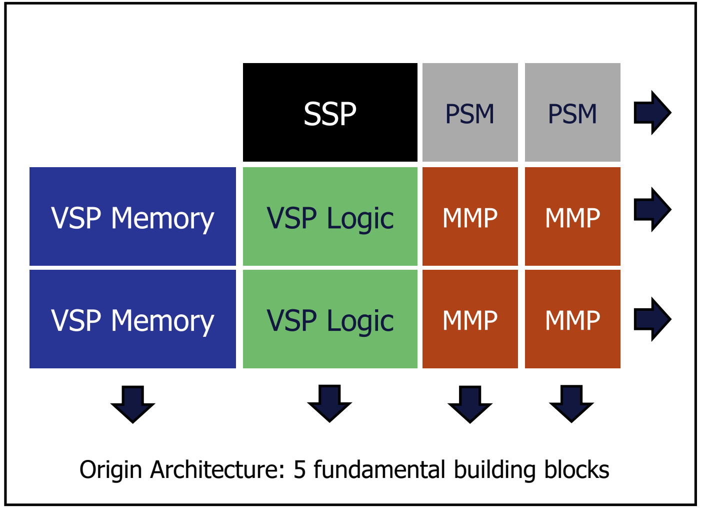
Origin Architecture: 5 fundamental building blocks

0.003 ~ 128 TOPS
Single core performance,
PetaOps with multi-core

18 TOPS/W
ResNet50 INT8 in TSMC 7nm @ 1GHz
No sparsity, compression, or pruning applied,
though supported

70-90%
Average sustained NPU utilization
across common networks

Architectural Building Blocks



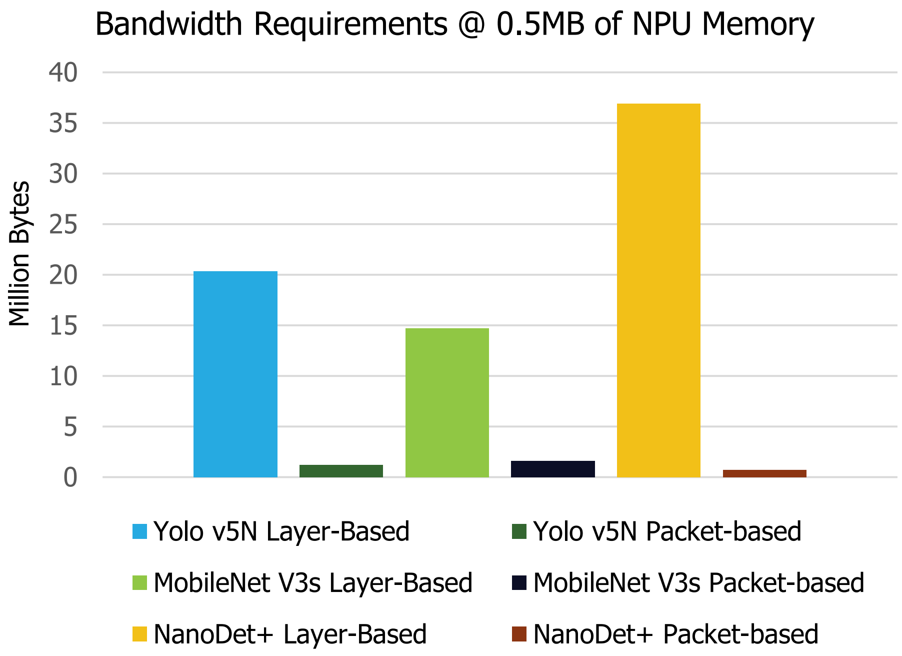
Origin Architecture: 5 fundamental building blocks

Decoupled building blocks, optimized for workload needs

- Scalable **matrix math processors** (MMP) perform weight multiplication.
- **Vector scalar processing** (VSP) logic handles memory access, data reshaping, and some vector operations.
- **VSP memory** provides storage for input, output, and intermediate activations.
- Accumulation buffering and quantization logic happen in **partial summation** (PSUM) blocks.
- Orchestrating operations is a single **sequence/scheduler processor** (SSP) working with compiler software for a unique packet-based sequencing approach.

Packets: Superior Memory Optimization

- Expedera’s packet-based architecture requires minimal external memory (weights only) for the networks shown
  - Higher throughput
  - Lower system power
  - Better privacy and security
- Uniformly spread-out bandwidth
  - Sustained utilization
  - Tolerance towards latency variations



Customer-provided, Field-proven Results

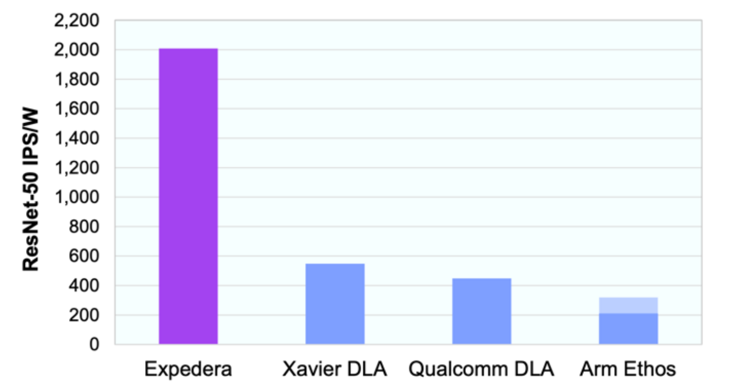
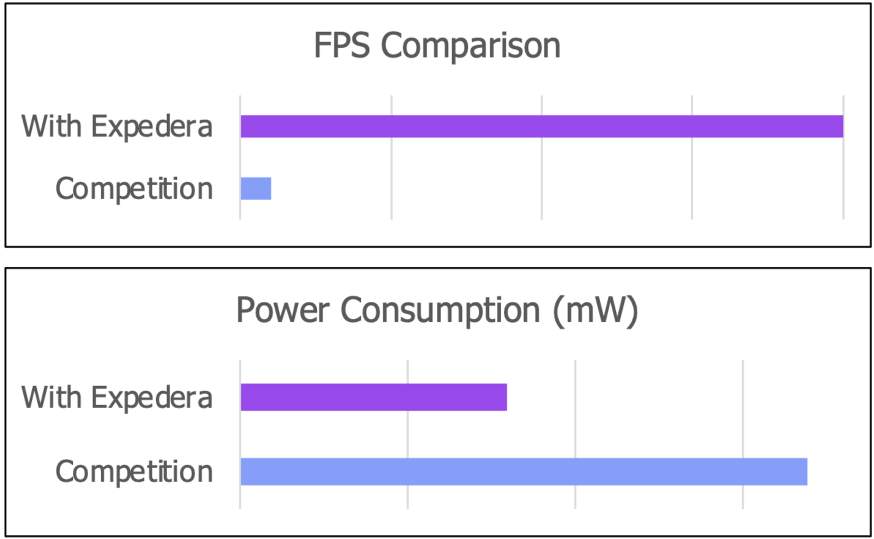


Figure 2. DLA-IP performance comparison. Expedera's performance and efficiency stand out against competitors'. The chart shows ResNet-50 v1.0 images per second (IPS) at best batch size versus typical power (W). (Source: vendors and The Linley Group estimates)

"Expedera Redefines AI Acceleration for the Edge"
- TechInsights (Linley) Microprocessor Report, April 2021



Customer-provided data, 4K video rate low light denoising:
20X faster throughput using less than half the power

Best Practices for Always-sensing Designs

Little NPUs achieve necessary performance within strict power and area budgets;
don’t settle for the big NPU

Memory management
Keep all data within the always-sensing subsystem, extending battery life while enhancing security & privacy

Resources

Summit & Alliance Resources

- Visit us at booth #319
- Alliance website
  - https://www.edge-ai-vision.com/companies/expedera/

Expedera Resources

- Company Website
  - http://www.expedera.com/
  - White papers, technical briefs, webinars, other
- Pre-silicon PPA Estimations
  - Want cycle-accurate PPA numbers for your use case(s) well before silicon?
  - info@expedera.com
- Contact us directly
  - info@expedera.com