Systems Drive Silicon:
How machine learning is reshaping chip-land

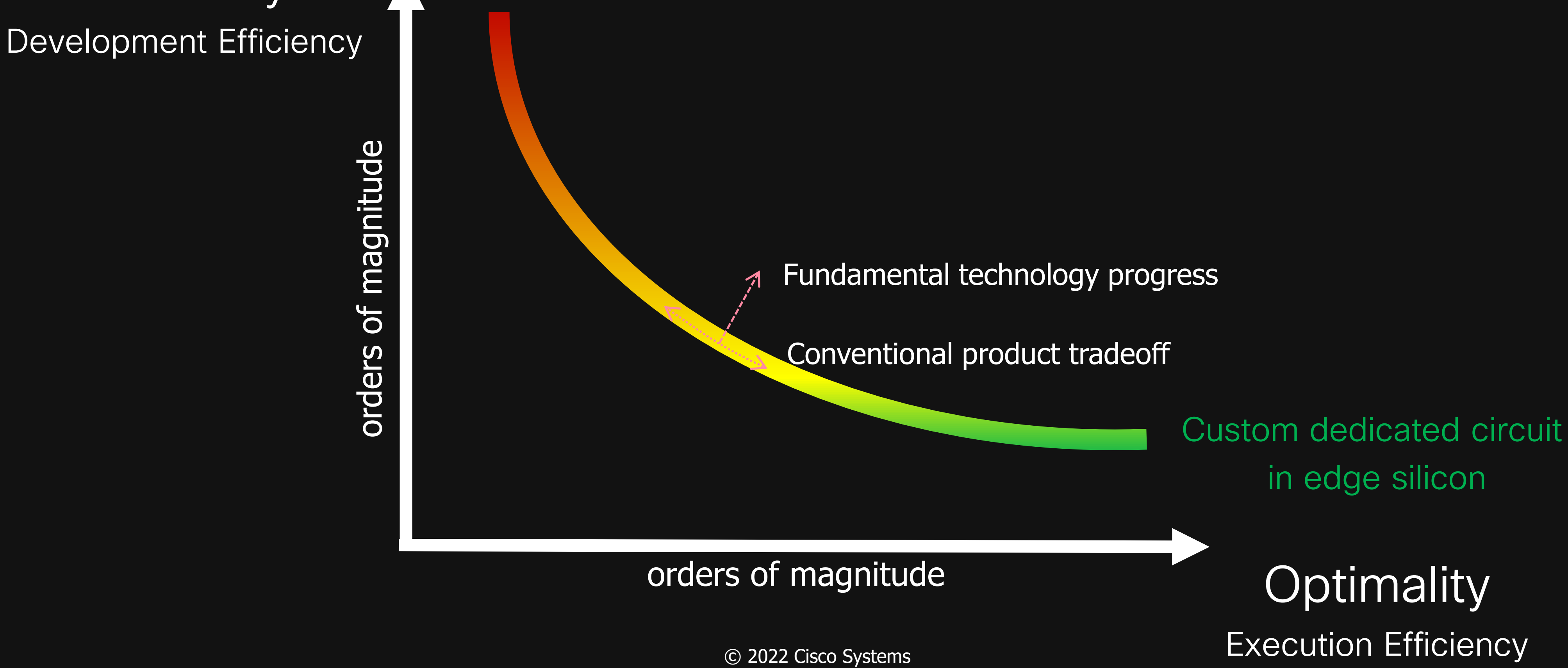
Dr. Chris Rowen
Cisco

October 2022

The Grand Tradeoff

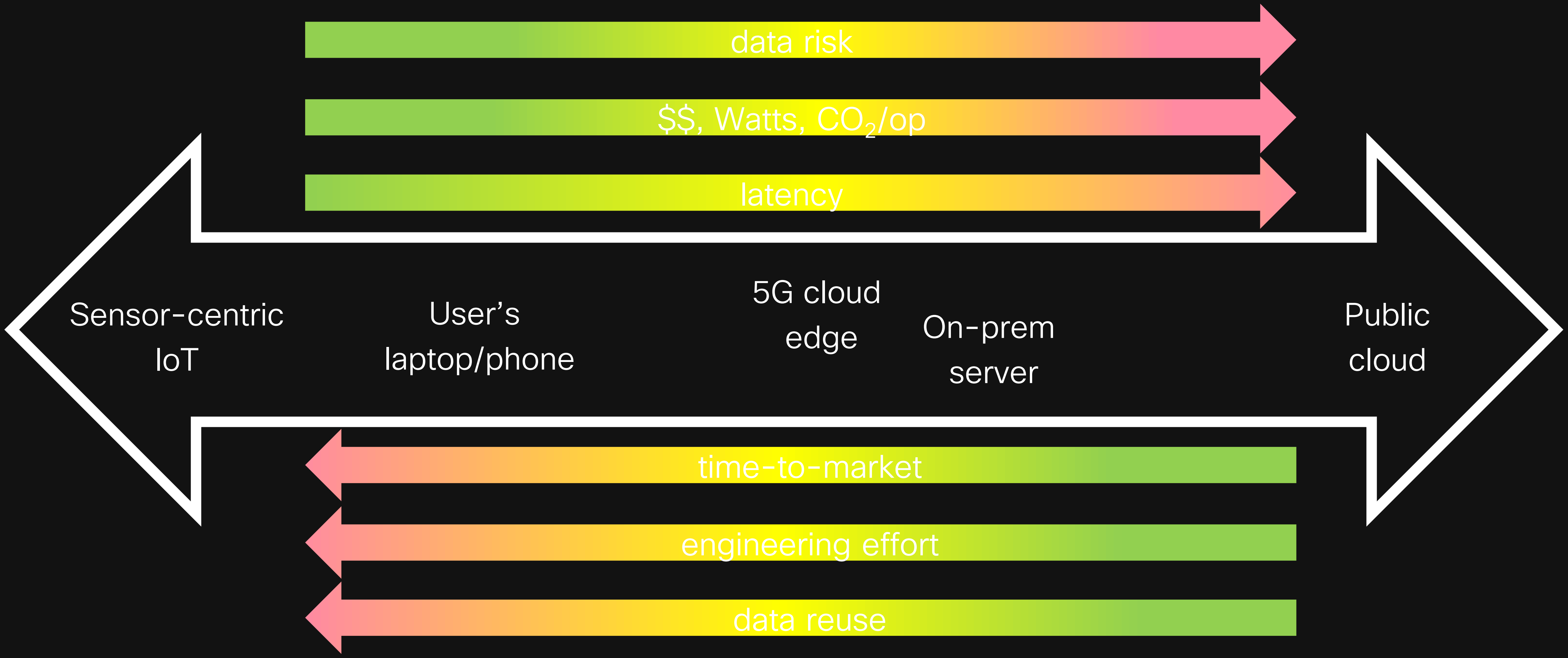
The most essential picture in tech

- How much more efficient must edge solutions be?
- What split of edge-cloud in hybrid systems?
- When is technology mature enough to freeze into silicon?



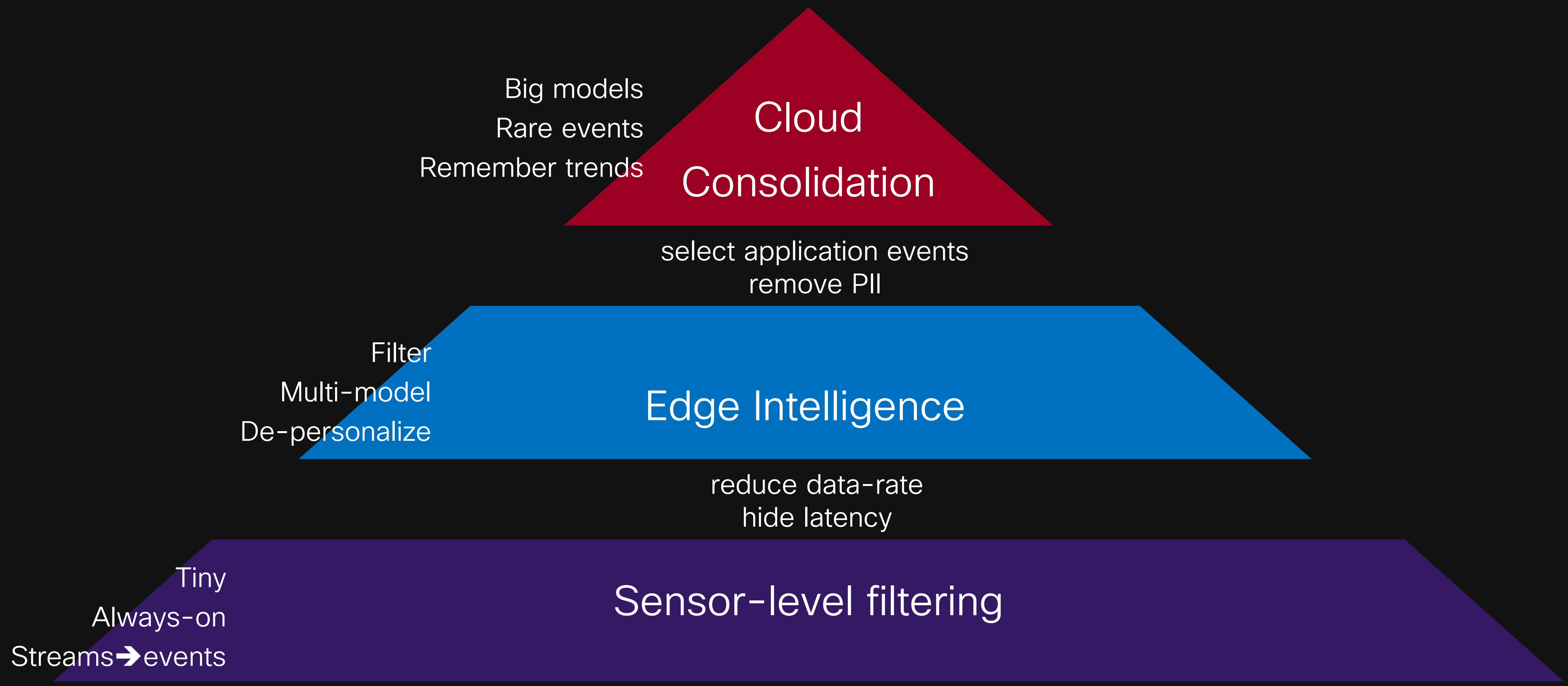
© 2022 Cisco Systems

Where to Compute



© 2022 Cisco Systems

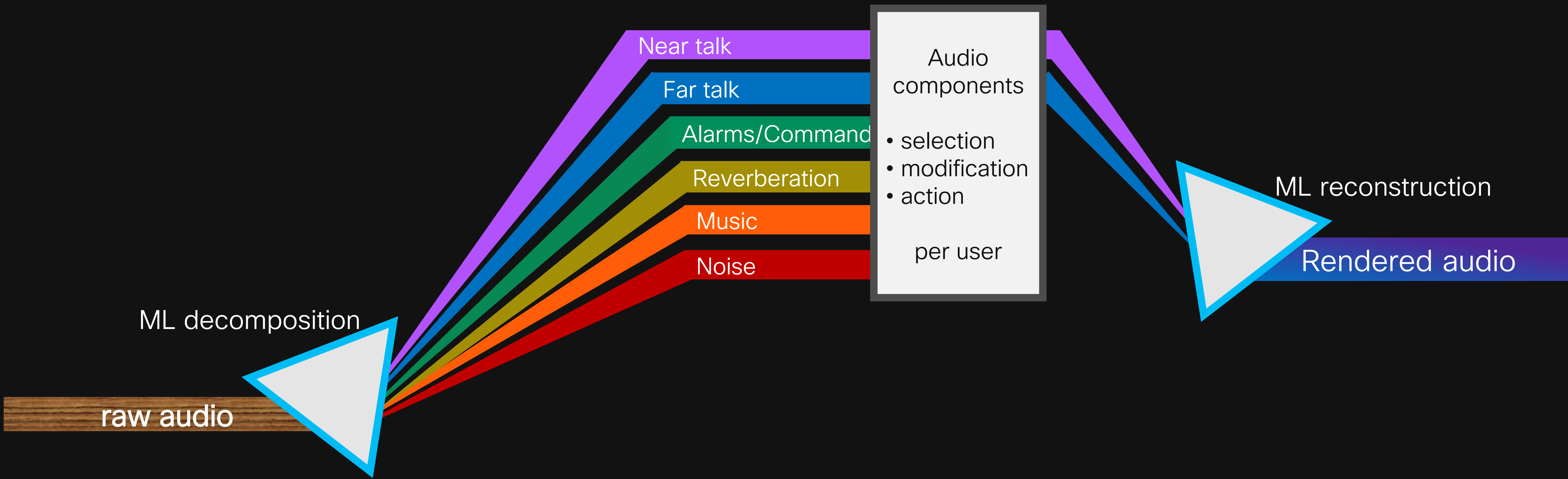
The Cognitive Hierarchy



© 2022 Cisco Systems

Rowen’s Prism

Decompose-Analyze-Reconstruct Audio



© 2022 Cisco Systems

The Audio Iceberg

The usual ML suspects:

- Noise reduction
- Speech-To-Text
- Text-To-Speech
- Talker ID
- Keyword trigger

ML below the surface

- Beam-forming
- Non-linear echo cancellation
- Voice activity detection
- Single talker isolation
- Background talker isolation
- Noise analysis/synthesis
- Voice cloning
- Prosody transfer
- Music identification/synthesis

- Packet loss concealment
- 3D source localization
- Source separation
- Talker-specific recognition
- Accent shifting
- Hybrid edge/cloud STT
- Tone/emotion analysis
- Equipment maintenance
- Underwater acoustic analysis

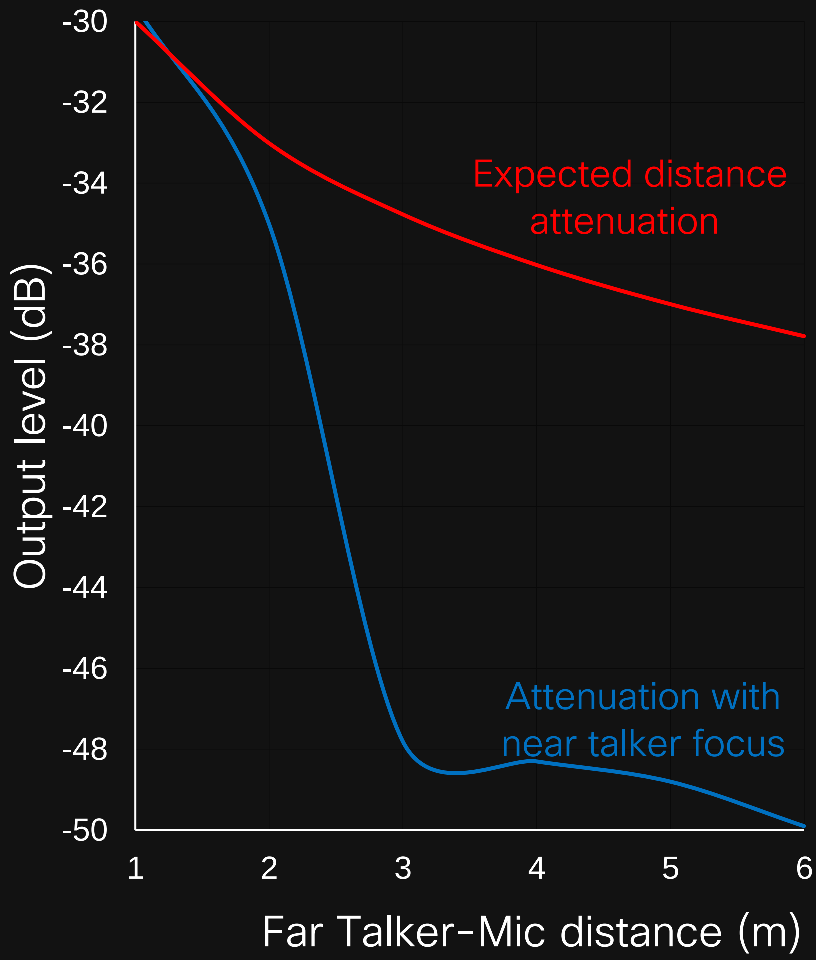
- Event classification – glass break, alarms, explosions
- Audio system diagnosis
- Source environment localization
- Health monitoring – Parkinson's, Alzheimers, autism, throat disease
- Language classification
- Dereverberation
- Pronunciation assessment
- Spoof detection

© 2022 Cisco Systems

Webex Audio Demo:
Noise Removal & Talker Selection

Noise removal (near-talker focus) and speech normalization use-cases

"Optimize for my voice"

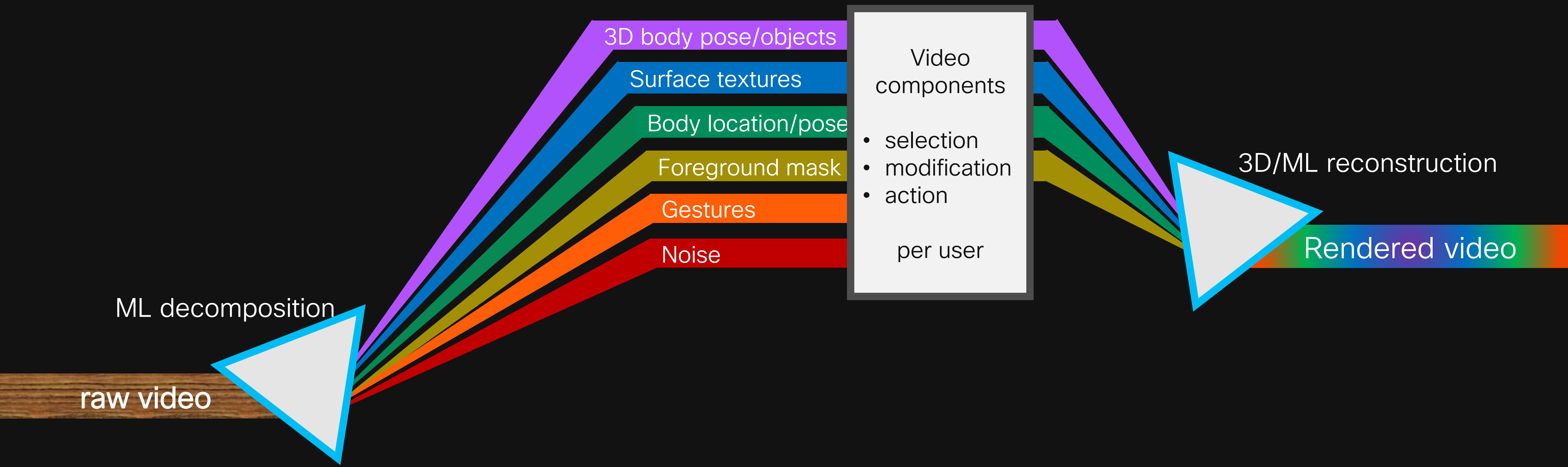


Live measurement on Android/Samsung Galaxy S8

© 2022 Cisco Systems

Rowen’s Prism

Decompose-Analyze-Reconstruct Video



© 2022 Cisco Systems

The Video Iceberg

The usual ML suspects:

- Object classification/localization
- Scene segmentation
- Face recognition

ML below the surface

- Gesture recognition
- 3D body pose
- 3D facial modeling
- Facial animation from audio
- Facial animation from text
- Liveness & spoofing detection
- Content-specific coding

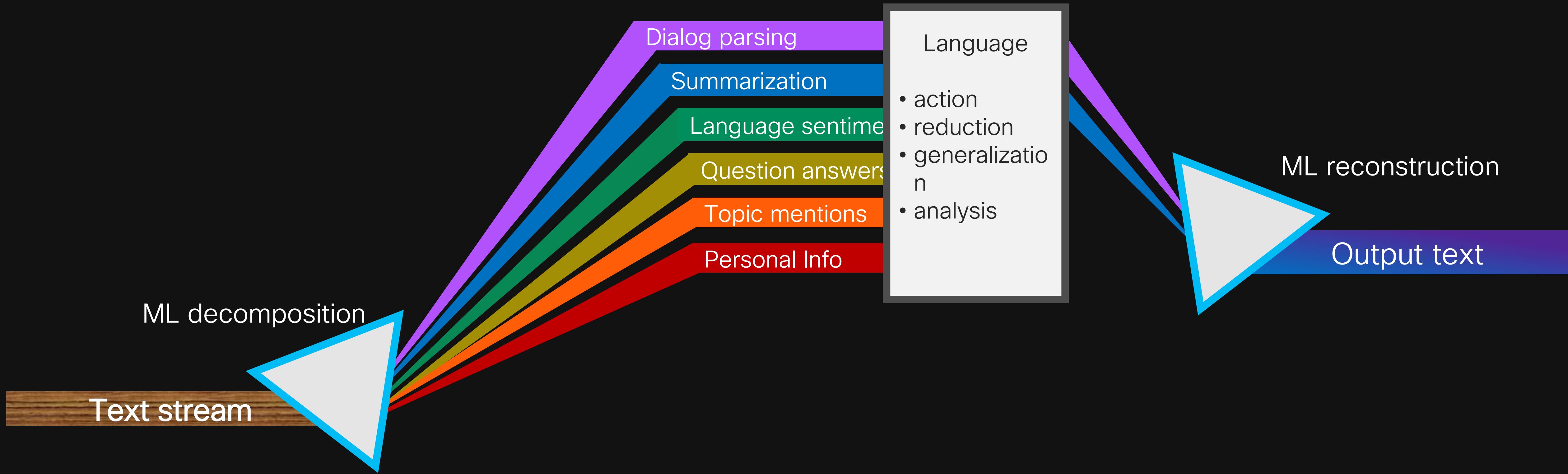
- Human super-resolution
- Sentiment analysis
- Demographic classification
- Face tracking
- Avatar generation
- User authentication
- Video content abridging

- Lighting/color correction
- Structure from motion
- Environmental assessment
- Visual search/matching
- People/object counting
- Health assessment from motion
- Content classification and digitization

© 2022 Cisco Systems

Rowen's Prism

Decompose-Analyze-Reconstruct Natural Language



webex

© 2022 Cisco Systems

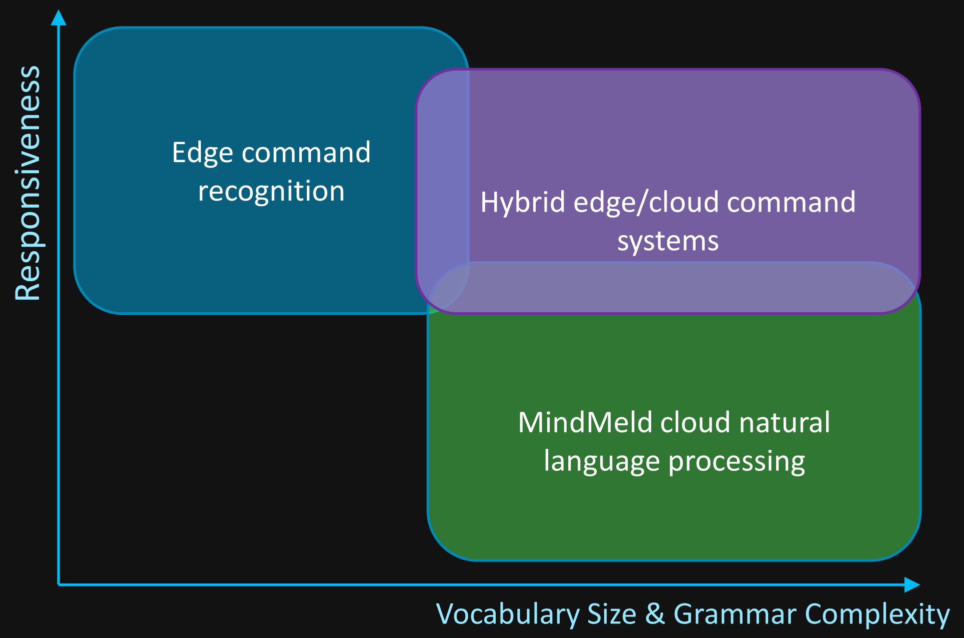
Example: Hybrid Edge/Cloud Speech Recognition
Cisco's Webex Assistant for Conference Rooms

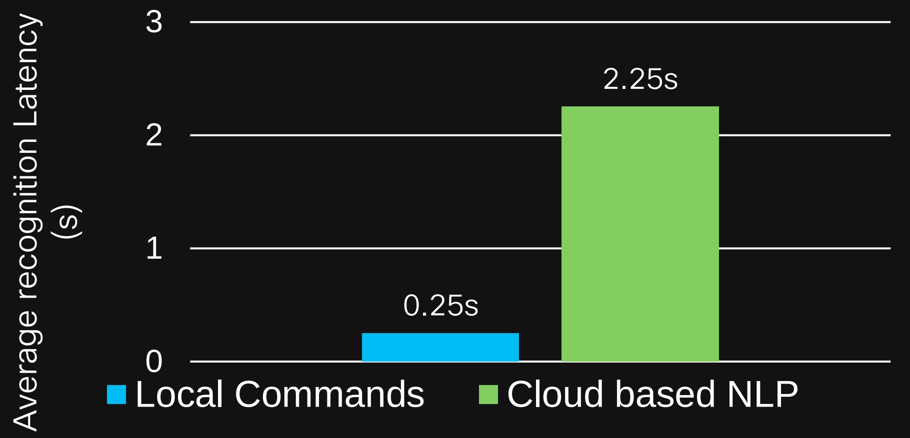
A large subset of commands executed locally on the endpoints, alongside the cloud NLP

Targeting ~1300 variations of 35 voice functions, which account for ~50% of usage

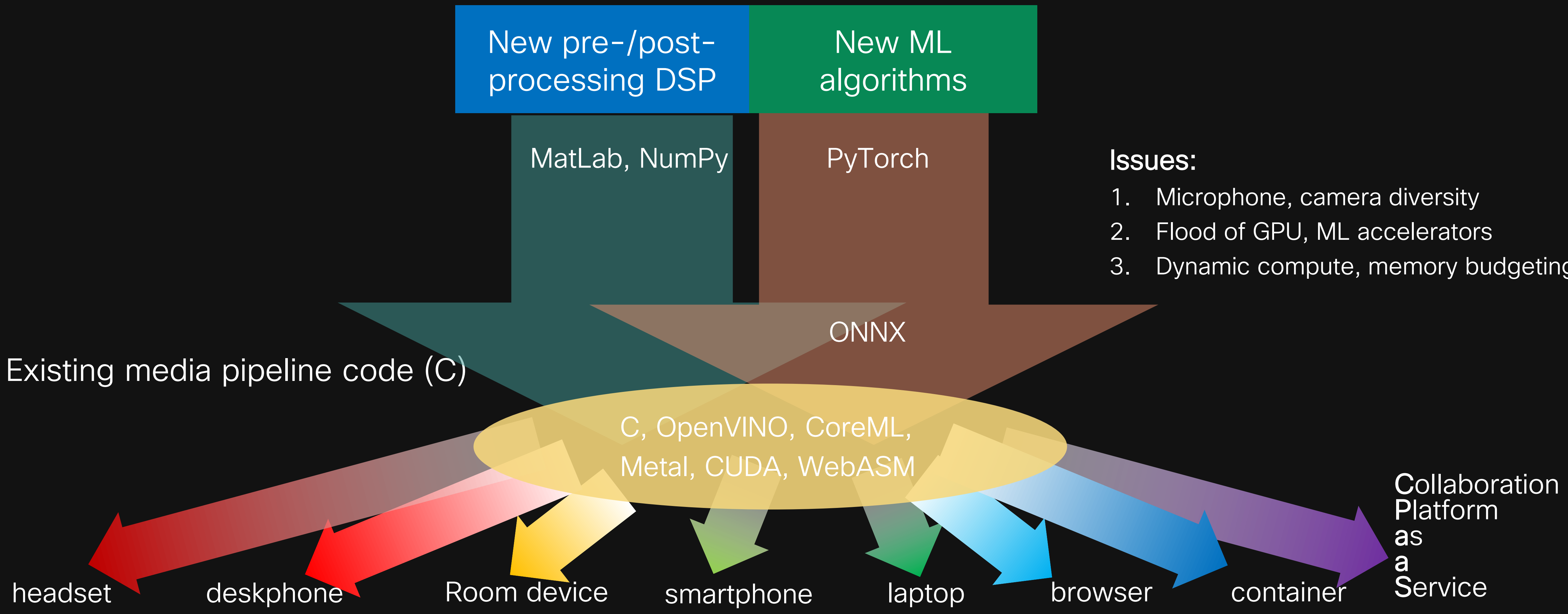
Order-of-magnitude improvement in recognitions tme

English, Spanish, French, German, Japanese, Portuguese, & Italian





Cisco/Webex Grand Challenge:
Heterogeneous Media ML Deployment

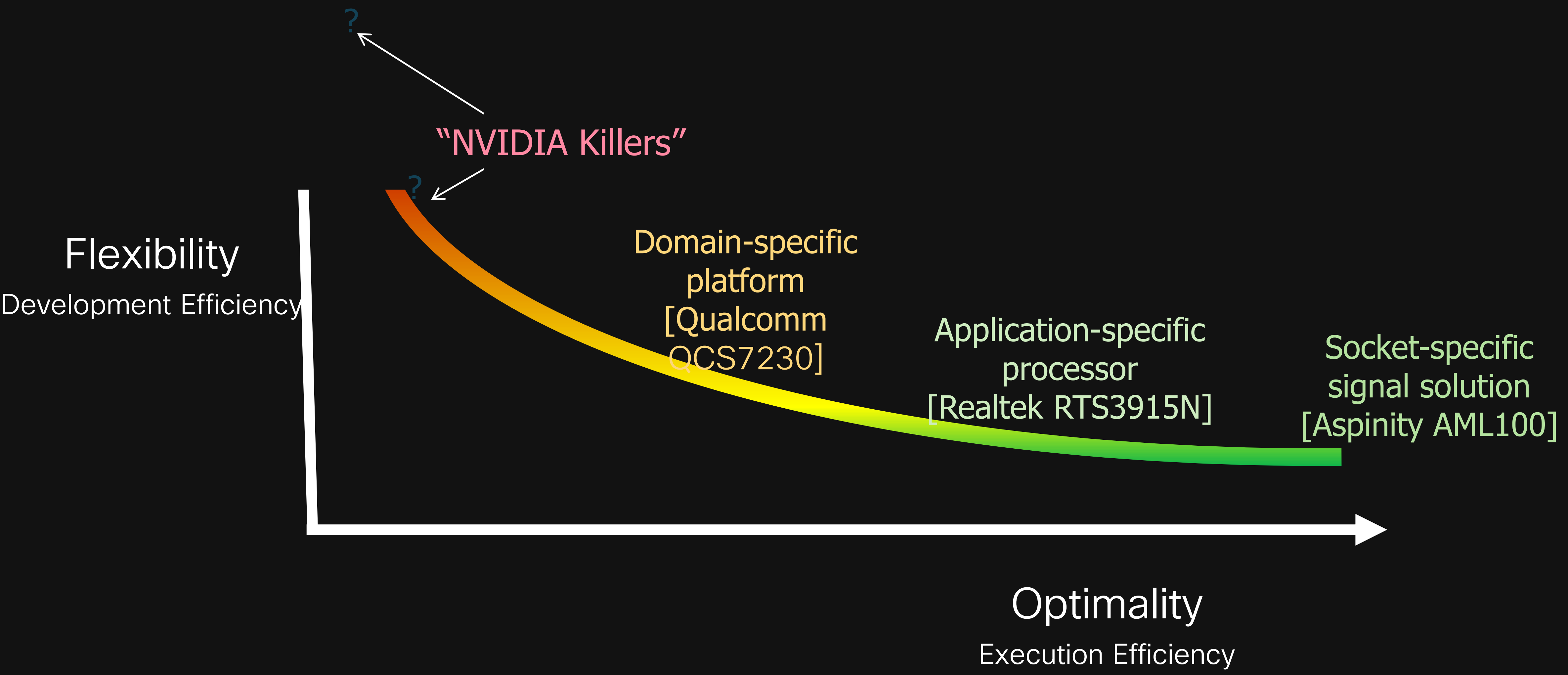


Issues:

1. Microphone, camera diversity
2. Flood of GPU, ML accelerators
3. Dynamic compute, memory budgeting

© 2022 Cisco Systems

Where Does ML Silicon Fit In?



© 2022 Cisco Systems

ML Acceleration in Silicon

Why? Multiply-add often dominates neural network compute

10^9 range ML power chip: Aspinity (~ 10s of μW) .. Cerebras (10s of kW)

Big factors in efficiency:

- Fraction of application in ML core network
- Utilization of compute units in real-world network structure
- Match of accelerator data-types to accuracy requirements (INT4 .. FP32)

ML Acceleration in Silicon

- To differentiate chip, exploit domain-specific characteristics:
  - Network sparsity
  - High compute:memory
  - Low-resolution data
  - In-memory compute
  - Analog multiply-add
  - Photonics
- The REAL bottleneck: mainstream network porting/optimization to architecture
  - Translator frameworks for complete network (ONNX)
  - Network layer libraries
  - Compiler: for ISA-integrated accelerations (vector processors)

Who is building AI silicon?

Established silicon vendors: Almost every CPU, DSP, GPU, MCU, SoC vendor has ML acceleration plan (plus some memory makers)

Lots of good ML accelerator IP – every major IP provider + 10 startups

Startups: 37	Sensor-specific	GP Edge	Server Optimized Inference	GP Cloud Training & Inference	Photonics
North America	Aistorm Analog Inference Aspinity Syntiant	Unthether AI Alif Semiconductor Blaize Gyrfalcon Kneron Mythic Perceive Rain Neuromorphics SiMa.ai Tetramem	Tenstorrent Groq Tachyum	Cerebras Ceremorphic Esperanto SambaNova Systems	Celestial AI Lightelligence Ligthmatter Luminous Computing
Europe	Innatera	GrAI Matter Labs Hailo		Graphcore	Lighton Salience labs
Asia*		Blue Ocean Cambricon Horizon Robotics Alpha ICs	Furiosa Neuchips		

Issues for Power in an ML World

Power is a system issue

- Choosing the right ML network architecture and training can make **100x** efficiency difference
- Silicon optimizations matter when algorithm is **mature.**
- Many ML applications are memory bound
- low power ➔ mass market ➔ responsible ML

Cutting through the hype

- Most application teams may port reluctantly to new ML hardware
- Accuracy impact of small data types (e.g. INT8) is hard to debug
- ML silicon impact likely quickest:
  - In video: compute bound
  - In embedded: chip is large fraction of power
  - In solution chips: efficiency advantage for whole data path, not just ML

The Next Five Years

Systems and Applications

1. Large Language Models
2. Inflexion point in ML compute cost on mainstream platforms
3. Slow but steady proliferation of ML know-how – basic tool for **every** software engineer
4. Closely-coupled hybrid system design: edge+cloud
5. Growth of "The Data Economy"

Silicon Design

1. Neural accelerators in every general-purpose chip
2. Pushing the envelope further on ultra high-end (memory bandwidth) and ultra low-end (analog compute)
3. Application-specific IC ➔ application-specific ML
4. Floating point remains dominant - NOT binarized, INT4, INT8
5. A few start-up successes

Some Resources

- My recent blogs on AI in collaboration: https://blog.webex.com/author/crowen/
- An earlier talk on audio/video ML startups: https://youtu.be/McFCQGO-SoQ
- Cisco's Responsible AI manifesto: https://blogs.cisco.com/security/introducing-cisco-responsible-ai-enhancing-technology-transparency-and-customer-trust
- Pushing ML to ultra-low-power – TinyML: https://www.tinyml.org/about/
- ONNX Tutorials: https://github.com/onnx/tutorials
- Audio ML with Python: https://opensource.com/article/19/9/audio-processing-machine-learning-python
- Video ML with Python: https://www.analyticsvidhya.com/blog/2018/09/deep-learning-video-classification-python/
- Recent funding in AI chip startups: https://www.wsj.com/articles/ai-chip-startups-pull-in-funding-as-they-navigate-supply-constraints-11647338402
- "95 AI chip startups": https://github.com/aolofsson/awesome-semiconductor-startups

© 2022 Cisco Systems