From Buildings to Living Systems

IoT to Edge-AI — Privacy-First Automation

Harsheeta Venkoba Rao — Founding Engineer, Gone.com | AI/ML Specialist | MSEE, UW Seattle

Why Now? The Shift to Edge Intelligence

15B+ IoT Devices

A global explosion of data sources.

Cloud Overload

Concerns over latency, cost, and privacy.

Edge-AI

Moving intelligence to the data source.

Efficiency Meets Ethics

Local decisions respect user privacy.

Factories & Smart Buildings — Systems of Systems

- **Subsystems:** HVAC, lighting, robotics, and security systems all act as independent "organs".
- **Connected via:** Open protocols like OPC-UA, MQTT, and REST APIs unify these subsystems.
- **System Intelligence:** Cross-domain coordination is what unlocks true collective intelligence and adaptation.

What Is Edge-AI, Exactly?

- **Local Inference:** Running AI models directly on or near the data source, not in a distant cloud.
- **Real-Time Action:** Enables immediate decisions (e.g., stopping a robot) without cloud round-trip latency.
- **Efficient Hardware:** Utilizes TinyML on specialized chips (TPUs, NPUs) for low-power operation.
- **Core Benefits:** Drastically improves power efficiency, reduces operational costs, and enhances data privacy.

Edge-AI in Action — Real Use Cases

Predictive Maintenance

Analyze vibration data locally to predict machine failure before it happens.

Smart HVAC & Lighting

Adjust climate and light based on real-time room occupancy and sunlight.

Defect Detection

Use local computer vision on assembly lines to find product flaws instantly.

Privacy-Preserving Security

Analyze security footage on-camera, sending only alerts (e.g., "person detected") not raw video.

Energy Optimization

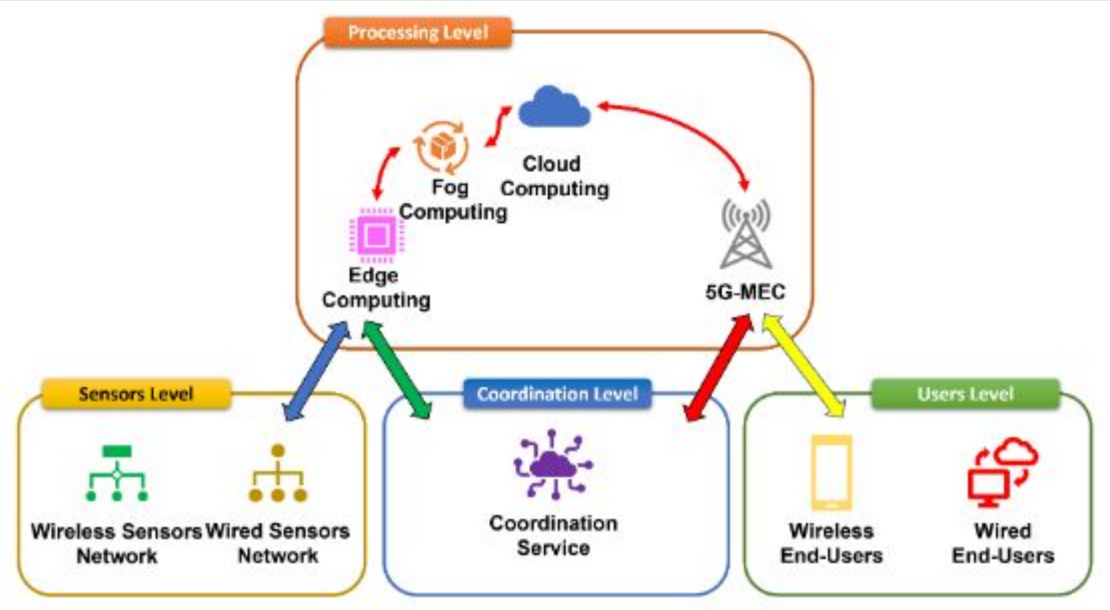
Intelligently manage building energy loads based on occupancy patterns and utility pricing.

The Privacy-First Blueprint

- **Local Processing:** Data stays on the device. Only insights (e.g., "room is empty") leave, minimizing exposure.
- **Federated Learning:** Improve a global AI model by sending only model updates from devices, not private user data.
- **Differential Privacy:** Add statistical "noise" to data summaries, making it impossible to re-identify individuals.
- **Regulatory Alignment:** Natively supports compliance with GDPR, NIST, and CCPA by minimizing data collection.

Designing Resilient Edge Architectures

- **Edge + Cloud:** Use edge nodes for real-time control and the cloud for global orchestration and analytics.
- **Offline Fallback:** Critical logic (like safety or comfort) must function even if network connectivity is lost.
- **Model Monitoring:** Continuously watch for "model drift" to ensure AI accuracy in changing environments.



Metrics That Matter

Metric Category	Engineering KPIs	Business KPIs	Sustainability KPIs
System Health	Uptime, Latency, Bandwidth Use	OEE, Asset Uptime, Yield	-
AI Model Health	Model Drift, Inference Time	(Tied to Business KPIs)	(Tied to Business KPIs)
Outcomes	Power Draw (mW)	Energy Cost Reduction	CO_2 per Building / per Unit
Human	-	Comfort Scores, Occupant Feedback	Air Quality (AQI)

Common Pitfalls to Avoid

Oversized Models

Deploying large, power-hungry cloud models on constrained edge hardware without optimization.

Environmental Variability

Ignoring sensor drift caused by real-world humidity, vibration, or temperature changes.

Lack of Integration

Creating intelligent subsystems (e.g., lighting) that don't communicate with other systems (e.g., HVAC).

Weak Endpoint Security

Failing to patch firmware or secure data, making edge devices an easy target for cyberattacks.

Optimizing AI for the Edge — Part 1: Compression

Pruning

Systematically removing redundant or non-essential neurons and connections from a neural network. This makes the model "sparse" and computationally lighter.

Quantization

Reducing the precision of the numbers used by the model (e.g., from 32-bit floats to 8-bit integers). This drastically cuts memory use and speeds up inference.

Optimizing AI for the Edge — Part 2: Knowledge Transfer

Transfer Learning

Adapting a large, pre-trained "global" model to a specific local task (e.g., a specific factory floor) using a small amount of new data.

Knowledge Distillation

Training a small, efficient "student" model to mimic the outputs of a large, complex "teacher" model. The student inherits wisdom at a fraction of the size.

Optimizing AI for the Edge — Part 3: Deployment & Monitoring

Real-Time Profiling

Continuously monitor inference latency, power consumption, and memory usage on the live device.

Continuous Retraining

Implement version control for models and use new data to retrain and redeploy improved versions.

A/B Testing

Safely roll out new models to a small subset of edge devices to test performance before full deployment.

Optimization Case Study: MobileNetV2 for Human Classification

- **Depthwise Separable Convolutions:** This is the core trick. Instead of one massive calculation, it splits the work into two, much smaller steps (Depthwise and Pointwise) to reduce computation by up to 9x.
- **Inverted Residuals:** A novel block structure that is "narrow -> wide -> narrow". It takes in a compressed (narrow) input, expands it for processing, and compresses it again for the output, maintaining efficiency.
- **Linear Bottlenecks:** The final "narrow" layer in a block uses a *linear* activation (no ReLU). This is critical, as it prevents the non-linear ReLU from destroying information in the low-dimensional, compressed space.

Optimizing AI for the Edge — Part 5: Efficient Fine-Tuning

LoRA (Low-Rank Adaptation)

Instead of re-training the *entire* model, LoRA freezes the original weights. It injects tiny, trainable "adapter" matrices into the layers. Only these small adapters (a tiny % of the total parameters) are updated, saving massive amounts of VRAM and time.

QLoRA (Quantized LoRA)

Takes LoRA a step further. The large, *frozen* base model is quantized (e.g., shrunk to 4-bit precision) to save VRAM. Then, the small LoRA adapters are trained on top of this highly compressed model. This allows fine-tuning enormous models on a single GPU.

The Road to Autonomic Environments

- Local Autonomy + Global Orchestration
- Continuous Learning and Self-Optimization
- Privacy-First by Default, Not as an Afterthought

Image Sources

https://pub.mdpi-res.com/sensors/sensors-21-03784/article_deploy/html/images/sensors-21-03784-g001.png?1622620990

Source: www.mdpi.com

https://www.mdpi.com/information/information-13-00089/article_deploy/html/images/information-13-00089-g001.png

Source: www.mdpi.com

https://www.couchbase.com/blog/wp-content/uploads/sites/1/2024/07/Couchbase-Mobile-Overview-1.png

Source: www.couchbase.com

https://static.vecteezy.com/system/resources/thumbnails/070/595/295/small/futuristic-digital-circuit-blueprint-abstract-technology-grid-with-glowing-blue-lines-on-dark-background-perfect-for-ai-systems-high-tech-interfaces-and-cyber-data-design-illustration-vector.jpg

Source: www.vecteezy.com

https://elements-resized.envatousercontent.com/elements-video-cover-images/files/0cdd0ebf-1c5c-4068-884a-321129ddd2f1/inline_image_preview.jpg?w=500&cf_fit=cover&q=85&format=auto&s=54b7b721ce71b9b7c46b18c6b135ec144bc23b66b2ce7485d175f23d566ad5b3

Source: elements.envato.com

Questions?

Final Reflection: Think Locally, Act Responsibly.

Harsheeta Venkoba Rao | harsheetamorey@gmail.com

LinkedIn: