Electronic Design

ALAN R. EARLS, Contributing Editor

Designing to Support Energy-Efficient Edge AI in Process Applications

AI is moving to the edge, and that's a good thing. But designers must prepare for voracious power demands and be ready with the right hardware and the right strategies to optimize efficiency.

Process control isn't only confined to traditional process industries. It's becoming synonymous with "smart" manufacturing, making sure wide-ranging and often-loosely coupled activities function smoothly together. This has involved strengthening the edge, reducing dependence on central, hierarchical controls, and moving toward more responsive, real-time, adaptive operations.

Edge computing has enabled this transition, delivering real-time monitoring and opening the door to algorithmic-based control at the edge.

Access to more data, faster, at the edge, has been a direct product of reduced latency. Decisions can now be made even quikcer. The trend toward edge computing—where as much processing as possible is conducted at or near data or process activity—is well established. But now, with more compute-intensive and power-hungry edge applications emerging, including generative AI, system designers need to reconsider what they're doing and how they can do it within the power constraints typical of the edge.

There's a natural synergy between edge computing and advanced AI. The general evolution of edge computing in smart process control will inevitably include AI as new system-on-chip solutions are optimized for this approach. By locating AI applications where the data is generated, edge computing can deliver better predictive analytics and control strategies, improving process-control systems.

But a challenge for all edge computing is finding adequate power. Some edge locations have ready access to AC line power, but others aren't so equipped. The problem becomes more acute with AI applications, which tend to be voracious consumers of power.

Getting the Edge AI-Ready: Power and Hardware Considerations

Start with an assessment of power already available or readily accessible. Consider how much AI power may be required and the feasibility of accommodating that with existing power sources such as wired on-site AC or DC power—or sources that could easily be added.

Battery power alone may be an option for some edge purposes. However, due to high AI power demands, it might be better considered as a supplement or backup to ensure operation even during peak power needs or in the event of an interruption.

Solar/battery may provide a longer-term option for edge operations that are outdoors and far from central power sources, but it probably can only be supplemental for most AI edge needs. More unusual power sources could include a Rotman lens-based rectifying antenna for millimeter-wave harvesting of unused 5G network energy. This has been applied for powering sensors, though it would likely be insufficient for all but the smallest AI processing needs. And, of course, simpler solutions that don't depend on variables like sunlight availability tend to be more reliable solutions.

Assessing Power Requirements for AI at the Edge

The amount of energy needed for edge AI varies with the sophistication and tasks being considered. One source cites a deployment of a GPU-based processor used for AI that consumes 20 to 30 W. The same node that performs processing might also need to power routing switches, sensors, etc., implying that it's better to plan for more power needs rather than less.

Monitoring power use through, for example, a battery-management system, is a good idea for both ongoing man-

agement and to meet any potential reporting requirements.

Power supplies must deliver efficiency and reliability while also, often, needing to be compact and ruggedized. The variable workloads provided by AI are an invitation to adaptive or intelligent power-management features to keep consumption as low as possible. Modularity and scalability can also be features to consider for supporting growing deployment or facility reconfigurations.

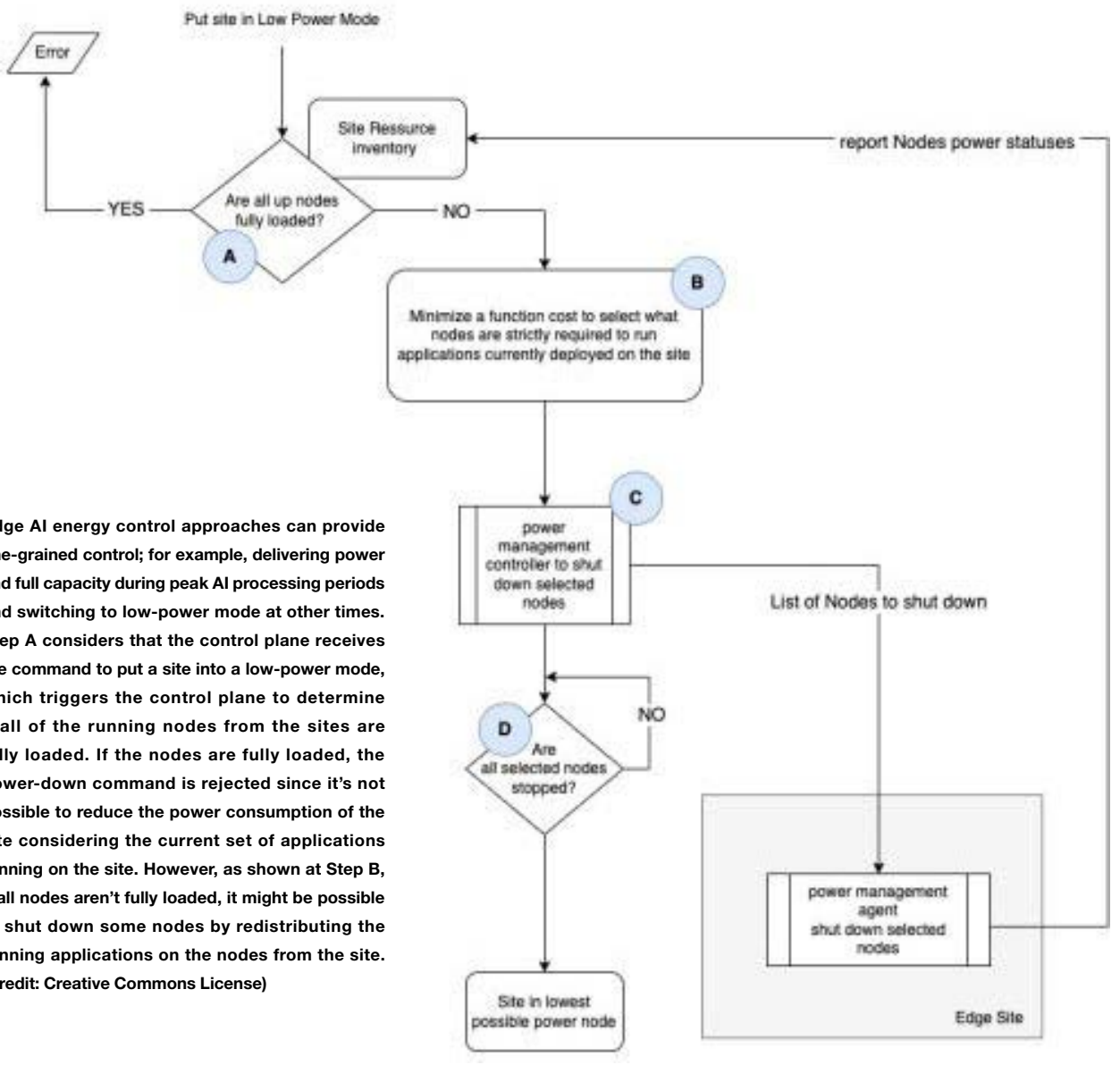
Deciding Between Cloud and On-Device Processing for Edge AI Apps

While the capabilities to "do" AI locally have grown, the power-management requirements may favor an alternative: the cloud. But that comes with caveats.

An obvious concern is latency. However, if a specific application can tolerate some latency, and where it makes more sense than processing in a traditional data center, the cloud is a viable option.

Processing in the cloud involves transmitting data and access to a cloud via the internet, which isn't without power needs, but this will typically be less than what's required to operate an AI application locally. Perhaps the biggest concerns are regarding security, though this is usually manageable, and latency, which may be harder to predict or control.

The authors of "Energy Management for Edge Sites" *(see References below)* note that energy savings may be possible when there are multiple compute nodes, and processing could potentially be throttled or even eliminated for some



Edge AI energy control approaches can provide fine-grained control; for example, delivering power and full capacity during peak AI processing periods and switching to low-power mode at other times. Step A considers that the control plane receives the command to put a site into a low-power mode, which triggers the control plane to determine if all of the running nodes from the sites are fully loaded. If the nodes are fully loaded, the power-down command is rejected since it's not possible to reduce the power consumption of the site considering the current set of applications running on the site. However, as shown at Step B, if all nodes aren't fully loaded, it might be possible to shut down some nodes by redistributing the running applications on the nodes from the site. (Credit: Creative Commons License)

periods of operation based on the specific use case. They also outline various schema for logically controlling such situations *(Fig. 1)* "by determining an optimal compromise between application availability and edge site power consumption without the use of specialized/bespoke hardware at edge sites."

Ensuring Reliable Data and Power Connectivity

While power and data movement are different domains, they should be considered at the same time. Power and data connectivity are both potential failure points; therefore, AI at the edge needs a fail-safe and/or a failover option—capturing state and delivering insights into the source and nature of a failure.

This should not be considered in a vacuum, but it should be relevant to the needs and expectations of the location or activity. What's the reliability posture of the operation as a whole? Are there backup power supplies or battery banks to ensure continuous operations? Alignment with the expectations of a whole facility or process is prudent to ensure relevance and avoid excessive cost.

And, of course, as a practical matter, wiring for power and data cabling can often be accomplished in coordination.

References

Energy Management for Edge Sites
A Practical Guide to Edge AI Power Efficiency