Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Elise Quevedo

Electricity and the future of AI

AI has created one of the largest infrastructure investment cycles in modern technology. However, another constraint is moving to the centre of the conversation: electricity.

The challenge is straightforward. More AI needs more computing capacity, and more computing capacity needs more power.

The electricity grid cannot expand at the same pace as software, and connecting large new data centres to the grid can take years.

In some US markets, developers are facing connection timelines of five to seven years or more.

This week, Emerald AI, Google and NVIDIA launched the AI Energy Management Alliance, or AEMA, bringing together companies across the AI, data centre, energy and utility sectors.

The objective is to advance flexible AI data centres that can adjust their electricity consumption in response to grid conditions.

From power consumer to grid participant

In the past, data centres were built with dependability in mind. To keep systems online, operators supplied power, ran computers continuously, and built in redundancy. Powerful AI is starting to change this equation.

In the end, not all computer tasks have to be completed at the same time. Certain work can be rescheduled, reduced in scope, or put on hold for a while. Batteries can offer additional assistance during periods of high demand, and on-site generating can augment grid power.

Software can coordinate these resources and respond to signals from utilities or grid operators. The result is a new kind of data centre. One that can be managed.

> The bigger opportunity is to use existing grid infrastructure more efficiently

This is the principle behind AEMA. Its framework focuses on measurable flexibility, including response speed, duration, predictability, and performance during grid emergencies.

The alliance is also exploring common technical requirements, operational data sharing, and interconnection approaches that recognise the benefits a flexible data centre can provide to the grid.

The bigger opportunity is to use existing grid infrastructure more efficiently.

The proof behind the concept

It is important to distinguish the new alliance from the broader concept of flexible data centres, because this is not starting from zero.

Google reported in March that it had incorporated 1 GW of demand-response capacity into long-term energy contracts with multiple US utilities.

The company says its data centres can shift or reduce portions of machine-learning workloads during periods when the grid needs additional capacity.

> The companies are now planning a nearly 100 MW power-flexible AI facility in Virginia with Digital Realty

Emerald AI and NVIDIA have also completed six showcases of flexible data-centre technology, according to Emerald AI CEO Varun Sivaram.

The companies are now planning a nearly 100 MW power-flexible AI facility in Virginia with Digital Realty, which is expected to provide a larger-scale test of the concept.

The industry needs operational proof, not another theoretical discussion about how AI could interact with the grid.

The 100 GW question

One of the most interesting claims surrounding AEMA is that greater flexibility could unlock up to 100 GW of additional capacity from the existing US power system.

This figure needs to be understood correctly. It is not 100 GW of new generation, nor does it represent capacity already unlocked.

It estimates how much additional data-centre capacity the grid could accommodate through greater flexibility in how large loads interact with the existing grid.

> AI will continue to require substantial amounts of additional electricity as deployments grow

Demand response cannot eliminate the need for new generation, transmission, or distribution infrastructure.

AI will continue to require substantial amounts of additional electricity as deployments grow.

What flexibility can do is reduce the infrastructure needed to serve peak demand and make better use of existing capacity.

The business case will determine the outcome

Flexible data centres require sophisticated software, controls, storage, and potentially on-site generation.

Operators also need to understand how much flexibility they can provide without affecting service-level requirements or the performance of critical AI workloads.

> If a data centre can show reliable flexibility and, as a result, receive faster access to grid capacity, the value becomes easier to quantify

If a data centre can show reliable flexibility and, as a result, receive faster access to grid capacity, the value becomes easier to quantify.

A developer could potentially reduce connection delays, avoid some infrastructure costs and improve utilisation of existing grid capacity.

Flexible loads are flexible because they can shift peak demand periods and potentially reduce the need to build infrastructure primarily for a small number of high-demand hours.

That could create a more balanced relationship between the data centre and the grid.

Collaboration is becoming infrastructure

The alliance brings together companies that traditionally occupy very different parts of the technology and energy ecosystem.

Anthropic is involved alongside utilities including National Grid and AES, power producers including NRG and Constellation, and companies working across energy management, storage, data centres, and grid technology.

NVIDIA brings the computing platform. Google brings extensive data-centre and demand-response experience. Emerald AI brings its energy-management technology.

Utilities understand grid requirements and reliability. Power producers understand generation. Regulators ultimately determine the framework in which all of these systems operate. No single company can solve this problem independently.

> AI is connecting sectors that historically operated in separate conversations

This is becoming true across the technology industry. AI is connecting sectors that historically operated in separate conversations.

Semiconductors are connected to power. Data centres are connected to construction. Optical networks are connected to AI clusters. Energy is connected to cloud infrastructure.

As I have mentioned in previous pieces, the infrastructure stack is becoming a single interconnected system. That makes collaboration an operational requirement.

Could AEMA become a global model?

Will AEMA become the first of many alliances focused on AI and energy, or will it remain primarily a US experiment?

The Middle East is making major investments in data centres and autonomous AI infrastructure, Asia is continuing to increase semiconductor and digital infrastructure capacity, and Europe is dealing with grid limitations and lengthy infrastructure development cycles.

Each market has a different energy system and regulatory environment. Still, how do we build the digital infrastructure required for the next decade without placing excessive pressure on the physical infrastructure supporting it?

Some regions will focus on new generation, others will invest heavily in transmission, storage, or nuclear power.

> Flexible computing will become another tool in the broader infrastructure equation

Flexible computing will become another tool in the broader infrastructure equation.

That is why I would not view AEMA simply as a coalition designed to help data centres obtain electricity faster.

Its more interesting role could be in establishing a new relationship between digital infrastructure and the power system.

The next phase of AI infrastructure

The next chapter of conversations should focus on efficiency. How efficiently can we use electricity? How efficiently can we use existing grid capacity? How quickly can infrastructure respond to changes in demand? How do we build enough capacity without unnecessarily increasing electricity costs for the communities hosting these facilities?

These questions create opportunities well beyond the traditional AI semiconductor market.

Power management, batteries, cooling, grid software, energy generation, transmission, networking, and data-centre architecture all become part of the AI infrastructure story

Power management, batteries, cooling, grid software, energy generation, transmission, networking, and data-centre architecture all become part of the AI infrastructure story. It broadens the value chain.

From my point of view, this is the type of conversation the technology industry needs to have. We cannot build the digital economy while treating energy and physical infrastructure as someone else's problem.

AI is forcing these industries into the same conversation, and I see that as a positive

development.

As this transition accelerates, collaboration will, as always, be crucial. No company, regardless of its size, controls the entire infrastructure stack required to support the next generation of AI.

Will AEMA become the first of many coalitions around the world? Or will it prove to be an interesting experiment that eventually loses momentum?

We do not know yet. What we do know is that the problem is too large to be addressed behind closed doors.

If we want a technological future capable of supporting this era's innovation, we need open conversations among technology companies, utilities, energy producers, governments, investors, and the communities where this infrastructure is being built. AEMA is one experiment in how to do that.