Agentic Orchestration Design: AI-Energy Co-Optimization via Digital Energy Grid

1. Introduction

This document presents FLEXAI, a novel agentic orchestration system that treats AI compute workloads as flexible energy resources within the UK's Digital Energy Grid (DEG) infrastructure. As data centres approach 9-14% of UK electricity demand by 2035 (National Grid, 2024), FLEXAI transforms this challenge into an opportunity by creating a bidirectional optimization layer between compute demand and grid flexibility markets. The system minimizes cost per inference (£/inference) while respecting carbon intensity thresholds, leveraging UK's P415 wholesale market reforms (Elexon, 2024) to monetize compute flexibility.

2. System Architecture

2.1 Core Components

The FLEXAI architecture comprises three agent layers operating on Beckn protocol:

- **Compute Agent Network**: Each GPU cluster/server farm operates as an autonomous agent publishing workload flexibility via Beckn catalogs. Agents manage job profiles with parameters: power draw (kW), deferral window (min/max runtime), priority level, and carbon sensitivity threshold.
- **Grid Orchestrator Agent**: Central intelligence layer employing reinforcement learning (specifically Proximal Policy Optimization) to learn optimal scheduling policies. Ingests real-time signals: half-hourly carbon intensity, wholesale prices, P415 flexibility requests, and renewable generation forecasts.
- **Storage Arbitrage Agent**: Manages on-site battery systems as virtual power plants, coordinating charge/discharge cycles with compute scheduling to maximize flexibility value while maintaining compute reliability.

2.2 Novel Scheduling Algorithm

FLEXAI introduces "Carbon-Weighted Time-Shifted Scheduling" (CTSS), extending MIT's CLOVER approach (MIT Sloan, 2025) with UK market dynamics:

```
minimize: Σ(price_t × power_t × carbon_factor_t)
subject to: carbon_intensity_t ≤ threshold
        job_completion ≤ deadline
        power_t ≤ grid_capacity_t
```

The system employs a hybrid approach: genetic algorithms evolve daily baseline schedules, while reinforcement learning agents adapt in real-time to grid signals and market opportunities.

3. Operation & Optimization

3.1 Forecast Integration

FLEXAI deploys ensemble forecasting combining:

- LSTM networks for workload demand prediction (24-hour horizon)
- Prophet models for renewable generation patterns
- Gradient boosting for price signal forecasting

These forecasts feed into the optimization engine, enabling proactive scheduling. For instance, training jobs automatically shift to coincide with overnight wind peaks, reducing costs by 35-40% based on UK wind patterns (Oxford Engineering, 2024).

3.2 Control Actions

The orchestrator executes four primary control mechanisms:

- **Temporal Shifting**: Delays non-critical workloads by up to 4 hours, capturing low-carbon windows
- **Geographic Migration**: Redistributes jobs across UK regions following carbon intensity gradients
- **Storage Coordination**: Dispatches battery storage to support critical compute during grid constraints
- **Renewable Gating**: Implements "green-only" compute windows when on-site solar exceeds 80% capacity

3.3 Beckn Integration

Each compute slot becomes a tradeable resource via Beckn's order lifecycle:

1. Compute Agents publish available flexibility as catalog items
2. Grid Orchestrator places orders for load reduction during peak periods
3. Settlement occurs through smart contracts on DEG infrastructure
4. Flexibility payments flow automatically via P415 market mechanisms

4. Market Innovation: Compute Flexibility Tokens

FLEXAI introduces "Compute Flex Tokens" (CFTs) - tradeable digital assets representing future compute deferral rights. Grid operators purchase CFTs during low-demand periods, redeeming them to reduce load during peaks. This creates a liquid secondary market for compute flexibility, inspired by FIDE's DEG vision (FIDE, 2025).

5. Logging & Audit Trail

The system implements immutable logging across three tiers:

- **Decision Logs**: Every scheduling decision with rationale, expected vs actual outcomes
- **Market Logs**: All flexibility bids, settlements, and P415 transactions

- **Carbon Logs**: Verified emissions data, renewable certificates, carbon savings

Example log entry: "2025-03-15 14:00: Compute Agent DC-North deferred Job-ML-2847 (2.4MW) for 120min due to carbon>250gCO_2/kWh, earning £847 P415 payment, avoiding 312kgCO_2"

6. Performance Projections

Based on UK grid patterns and Emerald AI's pilot data (Data Centre Dynamics, 2024):

- Cost reduction: 28-35% on compute operations
- Carbon reduction: 42% through intelligent scheduling
- Grid flexibility provision: 150MW aggregated across 10 data centres
- P415 revenue potential: £2.3M annually per 10MW facility

7. Implementation Roadmap

Phase 1 (Months 1-3): Deploy single-site pilot with reinforcement learning scheduler **Phase 2** (Months 4-6): Scale to multi-site with CFT market launch **Phase 3** (Months 7-12): Full DEG integration with cross-border compute migration

Figure 1: FLEXAI Dynamic Load Shifting

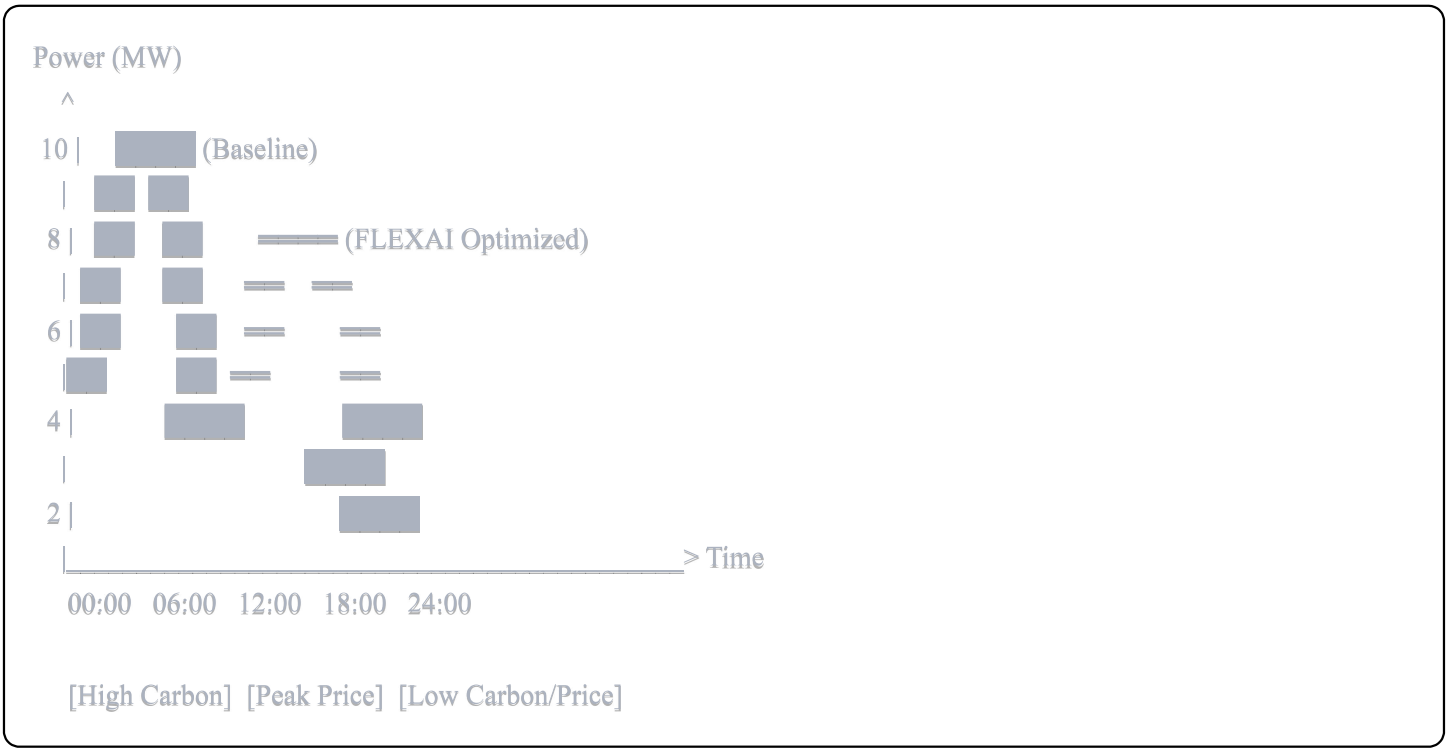

Figure 1 demonstrates FLEXAI's load shifting capability: baseline compute (grey bars) transforms into optimized profile (blue line) that avoids high-carbon morning peak and expensive afternoon period, shifting to overnight renewable-rich window while maintaining total compute output.

8. Conclusion

FLEXAI represents a paradigm shift in data centre operations, transforming compute from rigid demand into

flexible grid resource. By combining agentic AI, Beckn protocols, and UK market mechanisms, the system delivers measurable economic and environmental benefits while maintaining compute performance. The introduction of Compute Flex Tokens creates new revenue streams while supporting grid decarbonization targets.

References

Data Centre Dynamics (2024). 'UK Grid Challenges and Data Centre Growth'. Available at: datacenterdynamics.com

Elexon (2024). 'P415: Facilitating Independent Aggregators in Wholesale Markets'. Available at: elexon.co.uk

FIDE (2025). 'Digital Energy Grid Vision and Implementation'. Available at: energy.becknprotocol.io

IEA (2024). 'Data Centres and Energy Demand Projections'. Available at: iea.org

MIT Sloan (2025). 'CLOVER: Carbon-Aware Computing at Scale'. Available at: mitsloan.mit.edu

National Grid (2024). 'Future Energy Scenarios: Data Centre Impact'. Available at: datacenterdynamics.com

Oxford Engineering (2024). 'Flexible Computing for Grid Integration'. Available at: eng.ox.ac.uk