Benchmarking Hybrid Approaches to the Unit Commitment Problem

A hybrid quantum–classical decomposition workflow for industrial-scale UCP instances

JIJ Inc.,
ORCA Computing Limited,
National Quantum Computing Centre,
bp

June 2026

JIJ Inc.

Contents

Executive Summary

Purpose

This report benchmarks a hybrid quantum–classical approach to solving the Unit Commitment Problem (UCP) developed by JIJ, ORCA Computing, the UK National Quantum Computing Centre (NQCC), and bp. Reducing industrial energy costs is a priority within the UK's Industrial Strategy [1], particularly given that UK industrial electricity prices are estimated to be around 50% higher than the IEA average [2]. In this context, the report evaluates whether quantum computing could support large-scale energy optimisation and estimates when such approaches may become capable of solving problem instances relevant to bp's operations.

Key Findings

- The hybrid approach can solve instances of the UCP with **25,755 variables and 48,939 constraints on a real quantum processor.**
- The hybrid approach **can obtain higher-quality solutions than standard classical solvers**.
- The hybrid approach **can produce a more robust optimisation strategy** compared with classical day-ahead scheduling.
- Projections indicate that the hybrid approach, operating on quantum hardware that will be available later in 2026, could generate **higher-quality solutions faster than the classical state of the art for industrially relevant UCP instances**.

1 Introduction

1.1 UK National Context

Reducing energy costs is central to the UK's Industrial Strategy [1]. Industrial energy prices in the UK are currently around 50% higher than the IEA average, ranking the UK 6th among G7 countries [2]. Lowering these costs is imperative to maintaining the UK's attractiveness as a destination for investment. This imperative is especially true for attracting investment to the energy-intensive sectors the UK is targeting in its Industrial Strategy, such as advanced manufacturing, defence and life sciences. It will also be essential for the cost-effective operation of new AI data centres promised by the likes of NVIDIA [3] and Microsoft [4], following the UK–US Technology Prosperity Deal [5].

To address this challenge, the UK's Industrial Strategy and Clean Energy Industries Sector Plan emphasise the use of advanced technologies to optimise energy production and consumption and cut costs for industry and households [2]. Published shortly beforehand, the UK's National Quantum Missions [6] set out a parallel vision, including the development of UK-based quantum computers capable of delivering advantages over classical supercomputers across key sectors, including energy [7]. Together, these strategies create a timely opportunity to identify where quantum computing can most effectively improve efficiency and reduce energy costs.

In this context, JIJ, ORCA Computing, the UK's National Quantum Computing Centre (NQCC), and bp have collaborated to demonstrate a practical application of quantum computing in the energy sector by solving the unit commitment problem (UCP) using a hybrid decomposition algorithm executed on a photonic quantum system. This collaboration directly addresses this window of opportunity to chart a clear pathway to applying cutting-edge technology to a critical national challenge.

1.2 Unit Commitment Problem

The Unit Commitment Problem (UCP), outlined in Figure 1, is the challenge of scheduling power generator start-ups and shutdowns to meet electricity demand while minimising operational costs [8]. Industrial instances of the UCP involve upwards of hundreds of thousands of variables and constraints, making them computationally demanding for classical solvers. It requires long-term planning, with many constraints either tied to individual time intervals or spanning multiple time intervals. Figure 2 lists typical constraints and indicates their temporal scope.

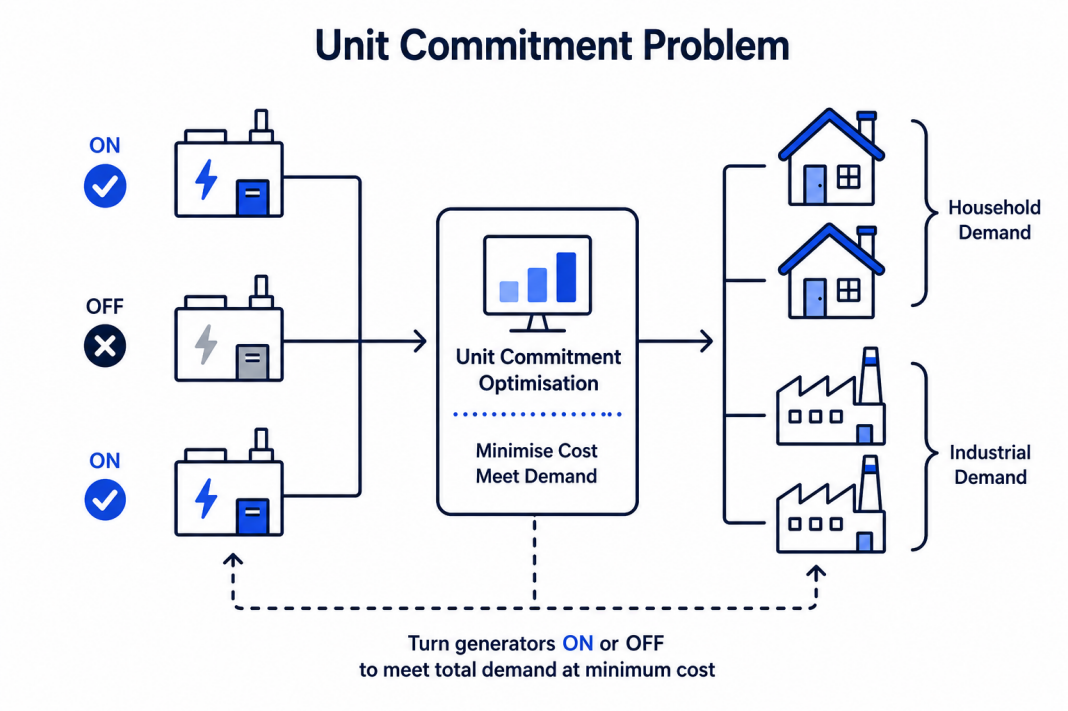

Figure 1: Schematic overview of the Unit Commitment Problem: a fleet of generators must be switched ON or OFF to meet the aggregate household and industrial demand at minimum cost.

The benchmarking undertaken in this project used the `unit_cal_7` dataset [9]. This dataset contains 25,755 variables and 48,939 constraints; it was selected by JIJ and verified as industrially relevant by bp's digital research and development team. Table 1 summarises the variables and parameters in the `unit_cal_7` dataset.

Our benchmarking work considers sub-instances of various sizes up to 25,755 variables. Though only the largest instance is of an industrially relevant scale, the results obtained at smaller sizes provide information about how the relative performance of the hybrid approach scales with increasing system size. This scaling analysis is presented in Section 2.1.

Typical constraints for this dataset include

start-up logic	$v_{g,t} \geq u_{g,t} - u_{g,t-1}$	$\forall g, \forall t$
output limits	$p_{g,t}^{\min} u_{g,t} \leq p_{g,t} \leq p_{g,t}^{\max} u_{g,t}$	$\forall g, \forall t$
power balance	$\sum_g p_{g,t} + s_t = d_t$	$\forall t$

Symbol	Kind	Type	Physical Meaning	Units
$u_{g,t}$	Variable	Binary	Generator commitment status (1 = generator g is ON at time t, 0 = OFF)	–
$p_{g,t}$	Variable	Continuous	Power output produced by generator g at time t	MW
$v_{g,t}$	Variable	Binary	Start-up indicator (1 if generator g starts up at time t)	–
$w_{g,t}$	Variable	Binary	Shutdown indicator (1 if generator g shuts down at time t)	–
s_t	Variable	Continuous	Load shedding variable representing unmet demand at time t	MW
d_t	Parameter	Continuous	Electricity demand at time t	MW
$c_{g,t}$	Parameter	Continuous	Operational cost variable associated with generator g production at time t	Currency
$p_{g,t}^{\max}$	Parameter	Continuous	Maximum generation level when generator g is on at time t	MW
$p_{g,t}^{\min}$	Parameter	Continuous	Minimum generation level when generator g is on at time t	MW
$r_{g,t}$	Parameter	Continuous	Spinning reserve provided by generator g at time t	MW
$ru_{g,t}$	Parameter	Continuous	Ramp-up variable controlling increase in generation of generator g at time t	MW
$rd_{g,t}$	Parameter	Continuous	Ramp-down variable controlling decrease in generation of generator g at time t	MW

Table 1: Variables and parameters in the `unit_cal_7` dataset.

The goal of the UCP is to minimise the total cost function

$$\min_{\text{variables}} \sum_t \sum_g \left[c_g(p_{g,t}) + c_g^{\text{start-up}} v_{g,t} + c_g^{\text{shut-down}} w_{g,t} \right],$$

subject to such constraints.

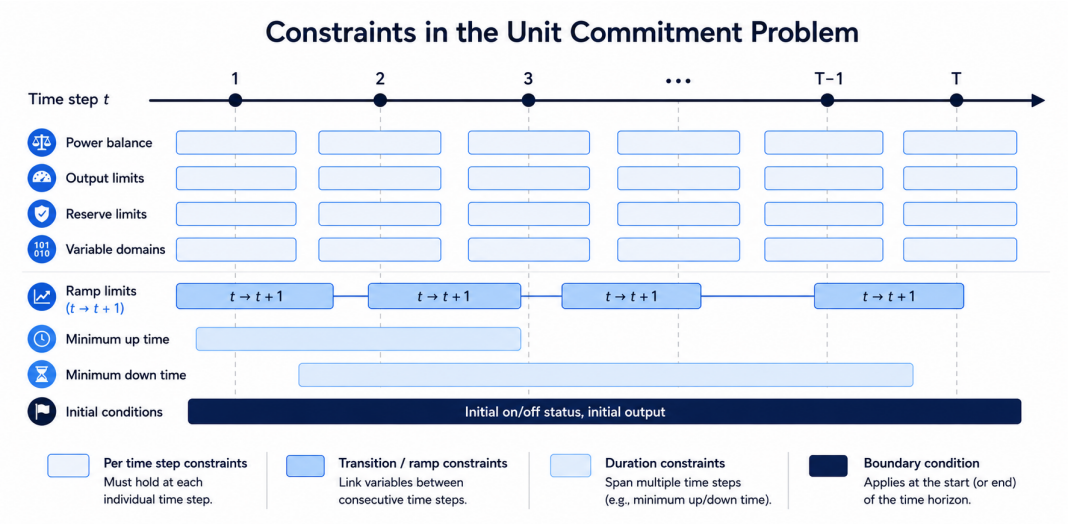

Figure 2: Constraints in the Unit Commitment Problem, classified by the temporal scope over which they act: per-time-step constraints, transition/ramp constraints, duration constraints, and boundary conditions.

With the definitions in Table 1, notice that an instance consisting of approximately 30 generators with data for each hour over a week (168 hours) would result in approximately 20,000 variables, making it a small but industrially relevant problem.

1.3 Hybrid Quantum–Classical Approach

The nature of the UCP encourages the use of decomposition methods in the solution. For references on some of the most popular methods, see the work of Montero and colleagues [8]. The proposed hybrid quantum–classical solution combines two of the common methods: the Dantzig–Wolfe decomposition [10] and the Benders decomposition [11, 12]. In particular, the Dantzig–Wolfe decomposition is used to divide the problem into distinct time slices, and the Benders decomposition is then applied to each time slice, as shown in Figure 3.

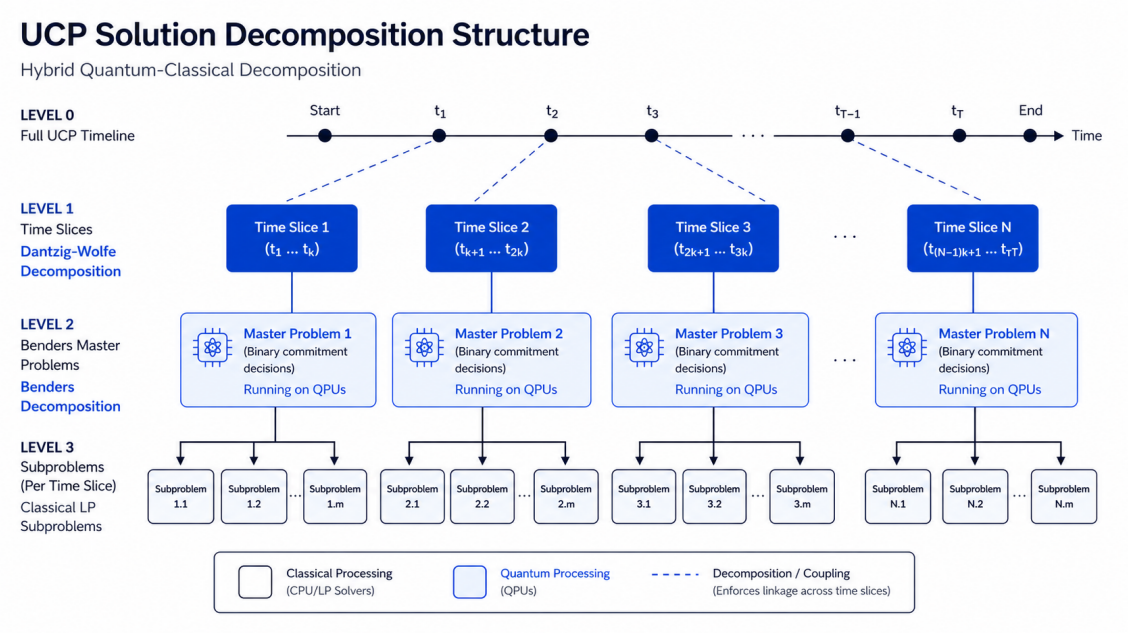

Figure 3: Solution overview. The hybrid workflow uses a Dantzig–Wolfe decomposition at the top level to split the full UCP timeline into time slices, and a Benders decomposition at the lower level to split each time slice into a binary master problem (executed on the QPU) and continuous classical subproblems.

The solution implements a staged workflow consisting of

1. Problem Formulation
2. First-Layer Decomposition (Dantzig–Wolfe decomposition)
3. Second-Layer Decomposition (Benders Decomposition)
4. QUBO Compilation
5. Optimisation Execution
6. Classical Post-Processing and Reconstruction

Problem Formulation. Subproblems from `unit_cal_7` are extracted with commitment, start-up, and shutdown decisions represented as binary variables, while power output levels remain continuous. The workflow uses JijModeling [13] to construct the model and generate structured optimisation instances. The framework separates model logic from instance data, allowing operational scenarios, demand profiles, decomposition strategies, and objective parameters to

be modified without altering the underlying mathematical formulation. This is particularly useful in industrial workflows, where models evolve continuously as operational constraints, reserve requirements, and planning assumptions change. Within the present workflow, JijModeling serves as the modelling layer through which optimisation logic can be developed, maintained, and adapted while preserving a consistent mathematical description of the underlying problem.

The resulting model is represented using Open Mathematical prograMming eXchange (OMMX) [14], an intermediate representation designed to preserve variables, constraints, objective functions, and metadata throughout the workflow. Whereas JijModeling is responsible for constructing the model, OMMX provides a stable representation of that model that can be passed between preprocessing, decomposition, compilation, benchmarking, and execution stages without repeated reformulation. This helps maintain traceability between outputs and the original problem definition throughout the hybrid workflow.

Prior to decomposition, JijZept presolve functions are applied to remove redundant variables and constraints and simplify the optimisation instance where possible.

First-Layer Decomposition. The OMMX instance generated in the previous stage is then processed using a Dantzig–Wolfe decomposition (DWD). This produces a classical Restricted Master Problem (RMP) that enforces global coupling constraints, such as power balance and reserve requirements, along with a set of subproblems that generate candidate schedules using reduced-cost information from the master problem.

As the optimisation instances are represented using OMMX, decomposition outputs remain in a common representation throughout subsequent stages in the workflow. This ensures that classical and quantum execution pathways operate on equivalent problem formulations, reducing implementation-specific differences between backends and improving the comparability of benchmarking results. It also provides a consistent framework for inspecting, reconstructing, and validating decomposition artefacts throughout the workflow.

Second-Layer Decomposition. At this stage, the subproblems defined in the Dantzig–Wolfe Decomposition are reduced via a Multi-Cut Benders Decomposition. Within this formulation, the Benders master problem contains the discrete binary commitment decisions and can be solved using the quantum optimiser running on ORCA's photonic quantum system, while the classical Benders subproblems evaluate dispatch feasibility and cost for candidate schedules. Dual information from the classical subproblems is used to generate multiple Benders cuts during each iteration. These cuts are added to the master problem to iteratively refine the solution space.

The output of this stage consists of subproblems whose discrete structure is suitable for quantum optimisation while preserving a mapping to the original UCP formulation through the OMMX representation.

QUBO Compilation. Following decomposition, subproblems are compiled into Quadratic Unconstrained Binary Optimisation (QUBO) form using Qamomile [14]. Qamomile provides an optimisation-oriented compilation layer that converts decomposed OMMX instances into executable QUBO Hamiltonians while preserving mappings between compiled variables and the original model. Maintaining these mappings is important because candidate solutions generated during quantum execution must subsequently be reconstructed and evaluated within the context of the original UCP formulation.

Constraints are incorporated through quadratic penalty terms, and penalty coefficients are selected to restrict the quantum solver to the feasible region of the search space. The resulting QUBO instances are then passed to the quantum execution layer.

Optimisation Execution. After the Benders Decomposition, the continuous variables are isolated and form a sub-problem that is efficiently solved by an LP solver to generate the mathematical feedback cuts for the quantum system.

Compiled QUBO instances are executed on ORCA Computing's PT-2 photonic quantum system based at the NQCC, using the Binary Bosonic Solver (BBS) [15]. The BBS operates as a hybrid variational algorithm. Rather than relying on a single quantum shot, it executes a feedback loop that iteratively tunes the interferometer's optical parameters to guide the probability distribution toward the QUBO's ground state. Each run produces a set of candidate solutions obtained through repeated sampling cycles. These solutions are stored in a candidate solution bank and passed to the classical orchestration layer for further evaluation.

Throughout this process, the workflow remains decoupled from backend-specific implementation details through the compilation layer provided by Qamomile. By preserving mappings between variables, compiled QUBO representations, and reconstructed solutions, candidate schedules can be evaluated consistently across different execution environments. This separation between optimisation logic and hardware execution simplifies benchmarking across heterogeneous compute platforms and helps ensure that performance comparisons remain focused on outcomes rather than software integration differences.

Classical Post-Processing and Reconstruction. The orchestration layer performs reconstruction by:

(a) stitching together optimisation outputs,

(b) evaluating boundary consistency between time intervals,

(c) applying local refinement where necessary, and

(d) verifying the resulting schedule against the original UCP formulation.

The preservation of model structure and variable mappings throughout the workflow allows candidate solutions generated during hybrid execution to be reconstructed and validated against the original optimisation model. Only schedules that pass classical feasibility validation are accepted as valid results.

2 Benchmarking

In this section, the performance of the hybrid quantum–classical approach described in Section 1.3 is compared against classical baselines. First, the scalability of the solution is tested against a classical solution which utilises the same decomposition strategy but employs a standard open-source solver (HiGHS) [16] in place of the Binary Bosonic Solver. This result highlights the importance of the quantum sub-routine in obtaining higher quality solutions. Second, the hybrid solution is placed in an operational setting, supporting a real-time scheduling strategy. It is compared against a standard day-ahead scheduling strategy supported by the HiGHS solver. This result demonstrates the hybrid solution's resilience to varying physical grid security constraints. Lastly, the hybrid solution's performance within a given timeframe is compared against various classical solutions, including Gurobi. This result demonstrates that the approach described is already capable of being competitive with state-of-the-art solutions and could outperform them by the end of 2026.

2.1 Scalability

This section details the scalability of the quantum–classical solution when compared against a classical solution utilising the same two-step decomposition strategy described above. The classical solution offloads the Benders master equation evaluations to HiGHS as opposed to the Binary Bosonic Solver provided by the ORCA photonic system. To test the scalability, a series of sub-problems of the `unit_cal_7` dataset were generated, the smallest consisted of 3,468 variables, and the largest of 17,340 variables. The value of the objective function evaluation for both solutions

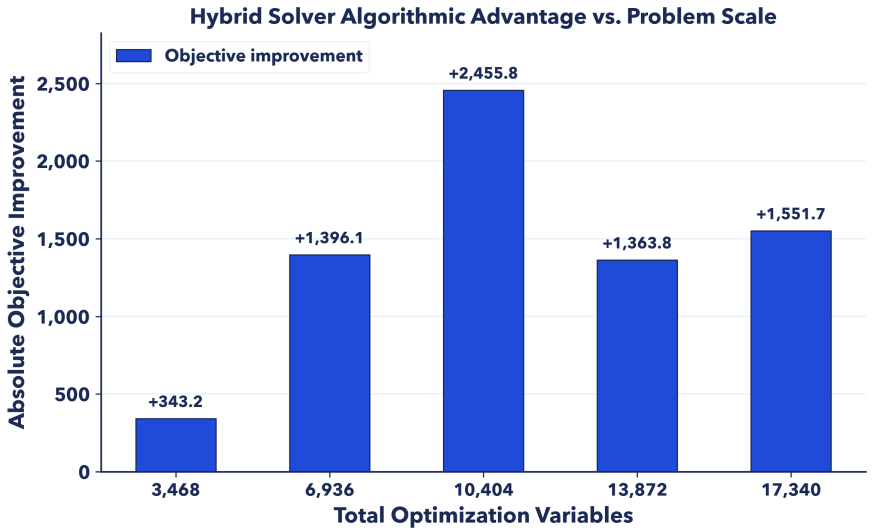

Figure 4: Scalability performance: absolute objective improvement of the hybrid solver over the classical decomposition baseline, plotted against problem size.

was gathered for each sub-problem, and the gap between the evaluations calculated. The results are shown in Table 2. As can be seen, the hybrid solver demonstrated strict superiority over the classical baseline across all tested problem scales. To observe the improvement gains as the problem size scales, the absolute objective improvement is plotted against the number of variables in Figure 4.

Total Variables	Baseline Objective	Hybrid Objective	Absolute Improvement	Relative Gap (%)	Block CVaR ($\alpha = 0.9$)
3,468	732,201.87	731,858.68	$+343.19$	-0.047%	0.00126
6,936	1,463,667.71	1,462,271.66	$+1{,}396.05$	-0.095%	0.00467
10,404	2,184,458.66	2,182,002.85	$+2{,}455.81$	-0.112%	0.00286
13,872	2,969,756.12	2,968,392.31	$+1{,}363.81$	-0.046%	0.00434
17,340	3,651,431.64	3,649,879.97	$+1{,}551.67$	-0.042%	0.00471

Table 2: Performance comparison across problem scales.

Note: A negative Relative Gap indicates the hybrid solution achieved a lower (more optimal) mathematical objective score than the baseline.

In utility-scale optimisation, objective score reductions directly translate to significant operational efficiency and resource conservation. The data confirm that the hybrid solution not only secures feasible continuous dispatch schedules, but also optimises discrete variables more effectively than the purely classical heuristic across the board.

2.2 Robustness

To validate the algorithmic robustness of the proposed hybrid quantum–classical approach, a sensitivity analysis was conducted on the 10,404-variable instance. The analysis evaluates the solution's resilience to varying physical grid security constraints, specifically focusing on the system's ability to undergo dynamic topological adaptation.

Real-Time Dynamic Adaptation vs. Static Day-Ahead Scheduling. As modern power grids integrate higher penetrations of intermittent renewable energy, operators frequently face unexpected intra-day fluctuations that require the sudden enforcement of stringent spinning reserve margins. To evaluate the hybrid solution's capability for dynamic adaptation in these scenarios, its performance was tested against a fixed, pre-computed classical baseline schedule,

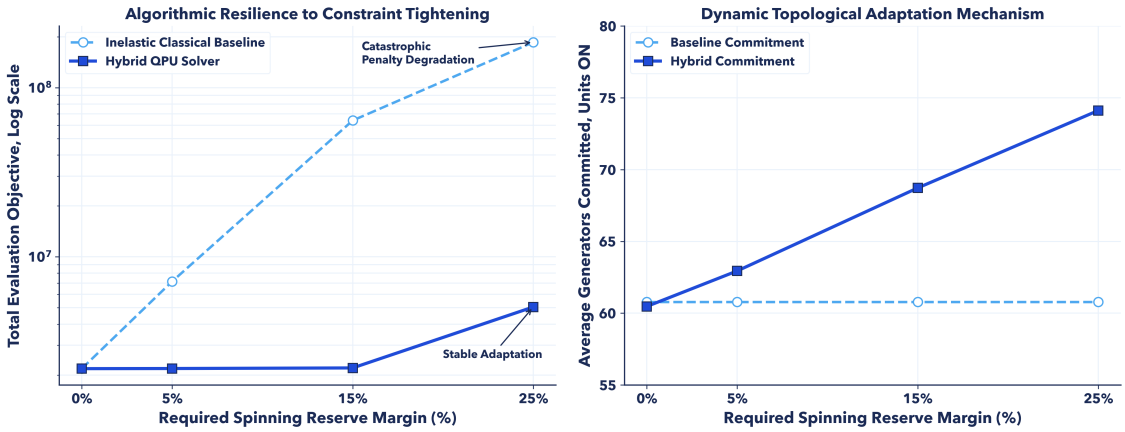

Figure 5: Robustness comparisons. *Left:* total evaluation objective (log scale) as a function of the required spinning reserve margin; the inelastic classical baseline degrades catastrophically while the hybrid QPU solver remains stable. *Right:* average number of committed generators as the reserve margin tightens; the hybrid commitment smoothly increases from $\sim$60 to $\sim$74 units, while the baseline commitment remains flat.

representing a rigid Day-Ahead commitment. Day-Ahead scheduling is used across industry, in some cases to align with market trading hours. Experimentation with Real-Time Dynamic Adaptation allowed the flexibility and elasticity of the hybrid solution to be evaluated. When graded against tightening security constraints, specifically, escalating spinning reserve requirements of 15% and 25%, this inelastic Day-Ahead baseline suffers from capacity shortfalls, as shown in Figure 5. Because its underlying operational topology is locked at an average commitment of approximately 60 online generators, it cannot adapt to the new constraints. This triggers grid failures and load-shedding penalties, degrading the total evaluation objective.

Conversely, the hybrid quantum–classical approach functions as a highly agile, real-time re-scheduler, demonstrating superior algorithmic resilience under these exact operational stress tests. By leveraging the QPU to sample a highly diverse combinatorial space, the hybrid solution's candidate banks maintain a rich set of ground states. This allows the classical assembler to rapidly pivot away from the Day-Ahead baseline when constraints change. As reserve requirements escalate, the quantum-assisted model identifies new feasible grid states and scales up the active generator fleet, smoothly increasing from 60 to over 74 committed units. By proactively bringing additional capacity online, the hybrid solver dynamically satisfies the 25% operational headroom requirement, avoiding the penalty regime while maintaining a stable, deeply optimised cost trajectory. This proves that quantum–hybrid methodologies can provide grid operators with a robust, adaptive scheduling tool that safely navigates extreme, real-time constraint environments where static operational schedules fail.

2.3 Solution Quality

Figure 6 compares time-to-quality performance on the largest-scale unit commitment instance with 25,755 variables. For consistent comparison and clearer visualisation, the y-axis reports normalised objective cost, with the best Gurobi solution serving as the ground-truth reference and set to 1.0. Under this metric, values closer to 1.0 indicate better solution quality.

We find that Gurobi provides a strong baseline performance, outperforming other solvers within a 5-minute timeframe. The open-source SCIP [17] and HiGHS [16] do not outperform Gurobi over this horizon, though the decomposition-based HiGHS workflow does show strong short-time behaviour which could be beneficial for time-constrained settings. The performance of these optimisers all saturates, with no significant improvements after 2 minutes.

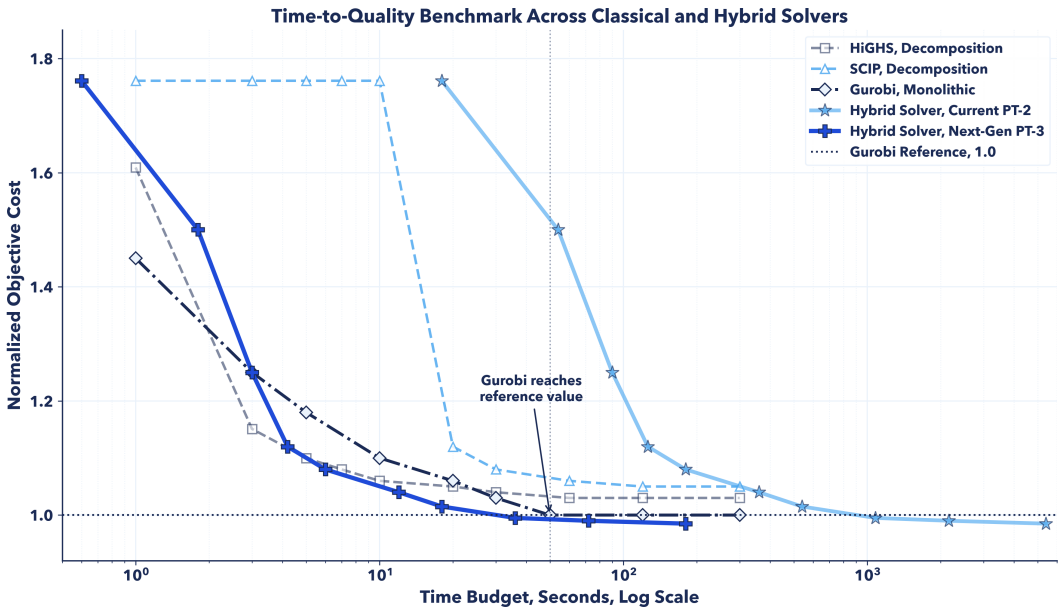

Figure 6: Solution quality comparison. Normalised objective cost versus time budget (log scale) on the ∼25,000-variable UCP instance. The Gurobi ground truth is set to 1.0. The hybrid solver on PT-2 keeps improving past where classical solvers saturate, and projections for PT-3 indicate it is expected to overtake Gurobi at shorter wall-clock times.

The best solutions overall are found by the hybrid solver using PT-2, albeit at longer timescales. Significantly, unlike the classical solvers that exhibit saturation behaviour and do not find improved solutions with more time, the solutions found by the hybrid solver with PT-2 continuously improve given more time. We note that this time is currently dominated by electronics and orchestration overheads, motivating projecting these results to the next-generation PT-3 system with $30\times$ faster wall-clock times. These projections to PT-3 indicate that the hybrid solver with PT-3 is expected to find better solutions faster than Gurobi. These results show that the hybrid approach using a quantum processor offers the strongest long-term scalability potential for large UCP instances when evaluated under a realistic time-budgeted benchmarking framework.

Key Insight: The solution quality from the hybrid quantum–classical workflow already matches state-of-the-art classical solvers on industrially relevant UCP instances, and projections to next-generation photonic hardware (PT-3) indicate it is set to compete with them on both solution quality and wall-clock time.

3 Outlook

This paper shows how quantum-enabled solutions are moving decisively from promise to practical application in tackling complex, large-scale industrial challenges. As demonstrated in Section 2.3, a hybrid quantum–classical approach developed by JIJ, ORCA, bp and the NQCC successfully solves Unit Commitment Problem (UCP) instances with 25,755 variables and 48,939 constraints on current-generation quantum hardware. Looking ahead, projections indicate that hybrid workflows running on quantum systems expected in 2026 could outperform leading classical methods, delivering higher-quality solutions in less time for industrially relevant UCP cases.

Section 2.2 further indicates that such hybrid approaches can be embedded within dynamic, near-real-time scheduling frameworks. This integration enhances solution resilience under conditions of system variability, including topology changes. The importance of such robustness is likely to grow as the share of intermittent renewable generation in-

creases.

Section 2.1 points to consistent performance gains as problem scale increases, underscoring the strong trajectory of both hardware and algorithmic development. These results reinforce confidence in the scalability of these approaches.

Taken together, the findings outline a compelling pathway to commercially viable quantum optimisation in the energy sector within the next generation of devices. This positions the UK to unlock advantages beyond classical supercomputing sooner than expected, supporting the ambitions of the UK's Industrial Strategy and Clean Energy Industries Sector Plan to harness advanced technologies for more efficient, cost-effective energy systems.

Contributors

Lead Authors

Dr Ross Grassie
Global Strategy & Operations Manager,
JIJ Europe Ltd.

Dr William Clements
Head of Applications and Software,
ORCA Computing Limited

Dr Manqoba Hlatshwayo,
Senior Quantum Applications Engineer,
National Quantum Computing Centre

Project Team

Dr Louis Chen
Global R& D Manager,
JIJ Europe Ltd.

James Fletcher
Head of Solutions Architecture,
ORCA Computing Limited

Alexander Makarovskiy
Quantum Machine Learning Scientist,
ORCA Computing Limited

Kavya Reddy
Project Manager,
ORCA Computing Limited

Manav Babel
Quantum Applications Engineer,
National Quantum Computing Centre

Claudia Perry
Digital Science Associate,
bp

Caitlin Rawcliffe
Digital Science Researcher,
bp

Acknowledgments

This research is based on benchmarking work carried out by JIJ, ORCA Computing, bp, and the UK National Quantum Computing Centre under the 2025 STFC Cross-Cluster Proof of Concept: UK National Quantum Computer Centre [NQCC200921].

A Hardware Scaling

To achieve industrial relevance, the hybrid approach must scale from small, decomposed problems to full industrial instances. Key scaling drivers include

- wall clock time of the quantum run,
- the number of variables optimised per quantum run,
- the number of decomposition tiles required, and
- the sampling complexity required for convergence.

As future generations of quantum hardware become available, the number of variables is expected to increase significantly while sampling overhead is expected to decrease, both of which will reduce the need for decomposition and improve time to solution. In this section, the scalability of the hybrid solution is analysed for larger problem sizes and how this will change with future hardware iterations.

A.1 ORCA Computing hardware: PT-2 and PT-3

The PT Series is ORCA Computing's product line of commercially available photonic quantum processors. These are non-universal quantum processors designed to support hybrid quantum-classical workloads in optimisation and machine learning. They combine fast speed of operations, scalability, and physical robustness for on-premises operation in a datacentre.

PT Series systems operate by producing, manipulating, and measuring quantum states of light known as "qumodes". While a qubit can be in either a 0 or 1 state, a qumode can contain either 0, 1, 2, or more photons. Compared to qubits implemented with matter-based quantum computing platforms, photonic systems offer long-range entanglement via the use of optical fibre delay lines, natural protection from several noise sources such as decoherence, and room-temperature operation. In the longer term, photonics also offers a scalable route to universal, error-corrected quantum computing in architectures involving multiple photonic chips connected by optical fibre.

A.1.1 PT-2

The ORCA PT-2 is the current generation PT Series. The work presented in this document used a PT-2 system hosted at the National Quantum Computing Centre. Its main specifications are:

- Maximum 48 qumodes
- Interference network consisting of two optical fibre delay lines
- Sampling rate of 300 ms per batch of 100 samples

A.1.2 PT-3

ORCA's next-generation PT-3 system, commercially available from mid-2026, will provide significantly reduced sampling overheads and the ability to solve larger problems. The target specification is 10 ms round-trip sample return time. This is a $30\times$ improvement over the PT-2's 300 ms. With these timings, the Binary Bosonic Solver (BBS) runs would require just $\sim$11 s of total quantum hardware sampling time, representing a saving of approximately 5 minutes per QUBO subproblem with the above settings.

Beyond latency improvements, the PT-3 offers substantially expanded hardware capabilities. The next-generation system is expected to achieve the following specifications:

- Maximum 128 qumodes

- Interference network consisting of three optical fibre delay lines
- Sampling rate of 10 ms per batch of 100 samples

The additional interference network structure will improve the performance of the BBS algorithm, and the increased qumode count allows for solving up to size-375 variable problems with ORCA's "tiling" technique from [15].

A.2 Runtime Projections

Figure 7 plots four different runtime projections for running BBS instances with the settings defined in the previous section:

- **NQCC PT-2:** Recorded average runtimes per QUBO subproblem with the above BBS settings, projected upwards across problem sizes.
- **PT-2 sampling:** The total minimum runtime for all sampling requests to a PT-2 during a run of BBS with the above settings. This is a theoretical lower bound assuming no additional classical runtime overheads.
- **PT-3 sampling:** The equivalent total minimum sampling runtime using PT-3 hardware specifications.

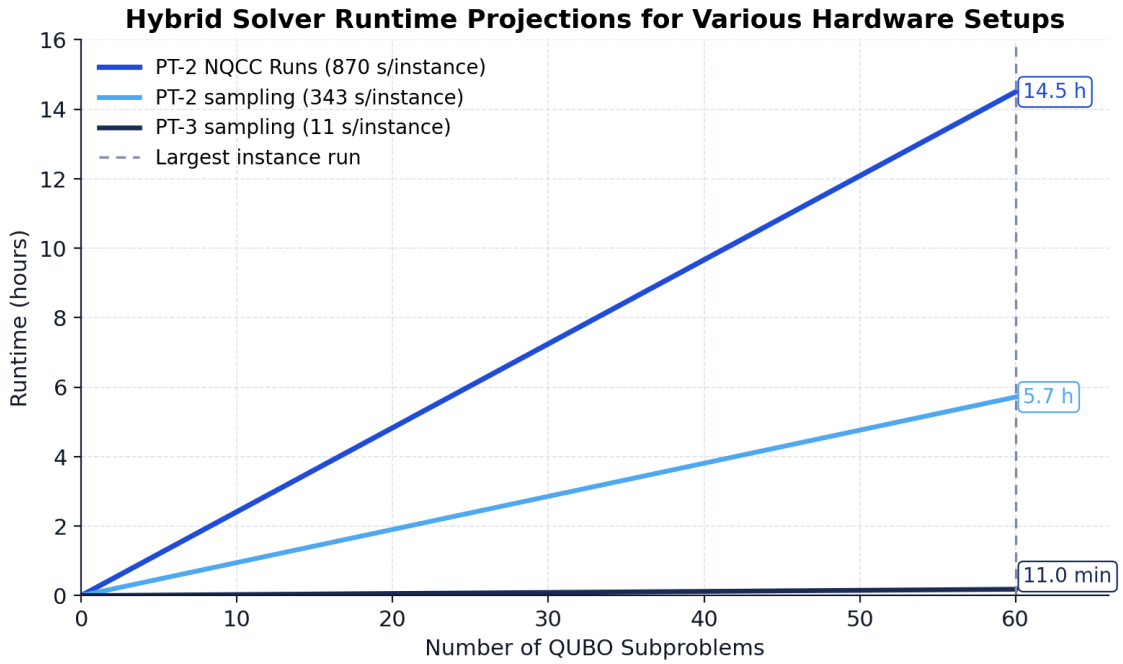

Figure 7: Hybrid solution runtime projections for various hardware setups. Recorded PT-2 NQCC runtimes are shown alongside theoretical PT-2 and PT-3 sampling-only lower bounds. The largest instance run is marked with a dashed vertical line; PT-3 reduces the projected runtime from ∼14.5 hours to ∼11 minutes.

The PT-2 and PT-3 sampling projections represent the minimum bound on total runtime achievable with the respective hardware, assuming a high-performance computing setup with a very low latency connection between classical and quantum compute, alongside powerful classical computing resources.

A clear difference is observable between the runtime measured using the NQCC PT-2 system and this minimum projected PT-2 runtime. This gap arises from several practical overheads not included in the projection, including the classical runtime of the BBS algorithm and communication latency between classical and quantum systems. In addition, the projection assumes a 300 ms minimum sampling time, which represents a lower bound. In practice, the effective sample call duration on hardware can be longer due to device-level operations and variability in experimental

execution, leading to runtimes that exceed this minimum.

A.3 Resource Limitations & Solution Quality

A.3.1 Qumode Count

The PT-2's limit of 48 qumodes means that large optimisation problems must be significantly decomposed into smaller QUBO subproblems before they can be executed on the hardware. This decomposition reduces the size of the optimisation landscape that can be explored in a single quantum run, which in turn limits the maximum solution quality outcomes we expect to attain. The smaller the subproblem, the less of the global problem structure is visible to the solver at any one time, which constrains the quality of the solutions it can find.

A.3.2 Loop Structure

The PT-2's interference structure, consisting of two optical fibre delay lines (or "loops"), limits the trainability of the BBS algorithm, which is an important factor in its optimisation performance. Future hardware generations will provide a larger loop structure such as the 1, 5, 25 length loops of the PT-3. The benefit of increasing to a three-loop structure for BBS has been studied previously in *A Binary Optimisation Algorithm for Near-Term Photonic Quantum Processors* (Makarovskiy et al., 2025), where longer 3-loop structures were shown to be beneficial for the BBS algorithm's performance.

A.3.3 Sampling Rates

Sampling rate directly affects both runtime and solution quality. A higher sampling rate allows more BBS updates to be performed within a fixed wall-clock budget, improving convergence and the quality of solutions found. Under current NQCC conditions, the PT-2's minimum 300 ms per-batch sampling overhead is the dominant runtime bottleneck. The PT-3's target rate of 10 ms per batch of 100 samples removes this as the primary constraint, allowing more sample-intensive runs to be considered for operationally relevant time windows.

A.4 Scalability Drivers

The primary constraint on scalability with the PT-2 is its 48-qumode limit, which caps QUBO subproblem size at 144 variables via tiling. This necessitates heavy decomposition of large problem instances, increasing the number of sequential quantum runs required and limiting the quality of solutions achievable per subproblem. The PT-2's two-loop structure also restricts BBS trainability, and the minimum 300 ms per-batch sampling overhead makes end-to-end runtimes impractical at industrial scale.

Future hardware generations are expected to address these scalability bottlenecks. The projected specifications of the PT-3 address each of these bottlenecks as discussed previously. Together, these changes make previously intractable problem sizes feasible, though deployment in a high-performance classical compute environment remains a prerequisite for achieving the projected runtime.

B Computing Environment

To ensure strict reproducibility and optimal execution performance, the test environment leverages an asymmetric hybrid computing paradigm. The computationally heavy presolving, mathematical decomposition, and QUBO compilation are executed locally on a high-performance classical machine. The resulting QUBO instances are then serialised and transmitted over a configured network to a remote photonic QPU, which acts as a specialised coprocessor for combinatorial ground-state sampling.

B.1 Classical Computing Environment (Host)

The classical node acts as the orchestrator of the decomposition pipeline, executing the ORCA/JIJ hybrid routine.

- **Hardware:** Apple M4 processor (arm64 architecture) equipped with 24 GB of unified physical memory. This unified memory architecture provides the high-bandwidth access required for rapid matrix transformations and the handling of large-scale QUBO coefficient matrices.
- **Operating System:** macOS 26.3 (Darwin kernel 25.3.0).
- **Programming Language:** Python 3.12.12, running in an isolated virtual environment to ensure execution efficiency and dependency stability.
- **Mathematical & Modelling Software:**
  - *Jij Suite:* `jijmodeling` (v1.14.1), `jijzepttools`, and `ommx`. These libraries handle the algebraic instantiation of the RMP and subproblems, slack binarisation, and automated QUBO compilation.
  - *Matrix Computation:* `numpy` (v2.3.5), utilised for high-speed tensor manipulations, QUBO COO-format processing, and robust median absolute calculations for dynamic penalty scaling.
  - *Data Visualisation:* `matplotlib` (v3.10.8), utilised for automated candidate generation heatmaps and baseline vs. hybrid cost comparison plotting.

B.2 Quantum Computing Environment (QPU)

The quantum subproblems are offloaded to an ORCA Computing photonic quantum solver, operating specifically on a Time-Bin Interferometer (TBI) architecture.

- **Hardware:** ORCA PT-2 Series Photonic Quantum Processing Unit.
- **Operation Mode:** Gaussian Boson Sampling (`gbs`) with photon number subtraction. The system maps the logical QUBO variables to the photonic qumodes and measures photon coincidence patterns to sample low-energy states of the target Hamiltonian.
- **Quantum SDK:** `ptseries` (v2.12.0), ORCA Computing's proprietary SDK containing the `BinaryBosonicSolver` (BBS), `create_tbi`, and `Logger` modules required to interface with the photonic hardware.
- **Solver Configurations:**
  - `modes`: 32
  - `tiles`: 3
  - `n_loops`: 1
  - `loop_lengths`: [1, 1]
  - `laser_power`: 1.0

  - `learning_rate`: 0.01 (with a flip learning rate of 0.1)
  - `updates`: 3
  - `n_samples`: 75 per iteration

- **Error Mitigation:** Hardware-level post-selection is enforced to filter out photon counts that may be empty due to loss: (`postselection=True`, `postselection_threshold=1`)

B.3 Network Configuration

Because the classical decomposition host (Apple M4) and the quantum solver (ORCA PT-2) are physically separated, the environment relies on a standard HTTP/REST network configuration to pass the instantiated QUBO matrices to the QPU.

- **Endpoint / API URI:** NQCC Host Networks.
- **Payload formatting:** The QUBO objects are scaled dynamically prior to transmission. The maximum absolute values of the linear and quadratic matrices are normalised to ensure the coefficients avoid precision washout over the network payload and fit safely within the physical limits of the TBI control electronics.

B.4 Execution Pipeline Setup

The testing environment is rigorously configured to run an exact apples-to-apples comparison between classical baseline heuristics and the ORCA PT-2 hybrid solver.

1. **Baseline Injection:** The test environment injects an exact baseline candidate per period (with `do_repair`, `do_trim`, and `do_polish` disabled) to ensure the Dynamic Programming (DP) assembler can perfectly reconstruct the classical baseline as a worst-case validation floor.
2. **Post-Processing:** Two forms of classical refinement algorithms are run locally on the M4 chip after quantum sampling:
   (a) *HLS (Horizon Local Search):* A 1-period coordinate descent over the horizon.
   (b) *HLS2 (Two-Period Local Search):* Targeted adjacent time-step swaps and 2-flips.
   (c) *Note on Test Controls:* Reserve margin shaping is strictly zeroed out during HLS/HLS2 repairs to ensure pure node-cost objective fidelity without heuristic bias during the evaluation phase: (`reserve_margin=0.0`, `reserve_weight=0.0`)

References

[1] Department for Business and Trade. “The UK’s Modern Industrial Strategy”. In: CP1451 (Nov. 2025).

[2] Department for Energy Security and Net Zero. “Clean energy industries sector plan: The UK’s modern industrial strategy”. In: (2025).

[3] NVIDIA Corporation. *NVIDIA Announces £2 Billion Investment in the United Kingdom AI Startup Ecosystem*. Sept. 2025. URL: https://nvidianews.nvidia.com/news/nvidia-announces-investment-in-the-united-kingdom-ai-startup-ecosystem.

[4] B. Smith. *Microsoft invests $30 billion in UK to power AI future - Microsoft On the Issues*. Sept. 2025. URL: https://blogs.microsoft.com/on-the-issues/2025/09/16/microsoft-30-billion-uk-ai-future/.

[5] Prime Minister’s Office. *Memorandum of Understanding between the Government of the United States of America and the Government of the United Kingdom of Great Britain and Northern Ireland regarding the Technology Prosperity Deal - GOV.UK*. Sept. 2025. URL: https://www.gov.uk/government/news/memorandum-of-understanding-between-the-government-of-the-united-states-of-america-and-the-government-of-the-united-kingdom-of-great-britain-and-north.

[6] Innovation Department for Science and Technology. *New Quantum Missions launched as Science Minister visits new advanced quantum lab*. Nov. 2023. URL: https://www.gov.uk/government/news/new-quantum-missions-launched-as-science-minister-visits-new-advanced-quantum-lab.

[7] Innovation Department for Science and Technology. *National Quantum Strategy Missions*. Dec. 2023. URL: https://www.gov.uk/government/publications/national-quantum-strategy/national-quantum-strategy-missions.

[8] Luis Montero, Antonio Bello, and Javier Reneses. “A Review on the Unit Commitment Problem: Approaches, Techniques, and Resolution Methods”. In: *Energies 2022, Vol. 15, Page 1296* 15.4 (Feb. 2022), p. 1296. ISSN: 1996-1073. DOI: 10.3390/EN15041296. URL: https://www.mdpi.com/1996-1073/15/4/1296/htm%20https://www.mdpi.com/1996-1073/15/4/1296.

[9] Zuse Institute Berlin. *unitcal_7 details*. URL: https://miplib.zib.de/instance_details_unitcal_7.html.

[10] George B. Dantzig and Philip Wolfe. “Decomposition Principle for Linear Programs”. In: *https://doi.org/10.1287/opre.8.1.101* 8 (1 Feb. 1960), pp. 101–111. ISSN: 0030-364X. DOI: 10.1287/OPRE.8.1.101. URL: %5Curl%7B/doi/pdf/10.1287/opre.8.1.101?download=true%7D.

[11] Ragheb Rahmaniani et al. “The Benders decomposition algorithm: A literature review”. In: *European Journal of Operational Research* 259 (3 June 2017), pp. 801–817. ISSN: 0377-2217. DOI: 10.1016/J.EJOR.2016.12.005. URL: https://www.sciencedirect.com/science/article/pii/S0377221716310244.

[12] John R. Birge and François V. Louveaux. “A multicut algorithm for two-stage stochastic linear programs”. In: *European Journal of Operational Research* 34 (3 Mar. 1988), pp. 384–392. ISSN: 0377-2217. DOI: 10.1016/0377-2217(88)90159-2. URL: https://www.sciencedirect.com/science/article/pii/0377221788901592.

[13] Jij Inc. *JijModeling: Mathematical modeling tool for optimization problem*. Python package, version 2.4.0. Released 23 April 2026; accessed 1 May 2026. 2026. URL: https://pypi.org/project/jijmodeling/.

[14] Wei-Hao Huang et al. “Qamomile: A Cross-SDK Bridge for Quantum Optimization”. In: *2025 IEEE International Conference on Quantum Computing and Engineering (QCE)*. Vol. 02. 2025, pp. 516–517. DOI: 10.1109/QCE65121.2025.10423.

[15] Alexander Makarovskiy et al. *A Binary Optimisation Algorithm for Near-Term Photonic Quantum Processors*. arXiv preprint. 2025. arXiv: 2510.08274. URL: https://arxiv.org/abs/2510.08274.

[16] Q Huangfu and J A J Hall. “Parallelizing the dual revised simplex method”. In: *Math. Prog. Comp* 10 (2018), pp. 119–142. DOI: 10.1007/s12532-017-0130-5. URL: https://doi.org/10.1007/s12532-017-0130-5.

[17] Tobias Achterberg. “SCIP: Solving Constraint Integer Programs”. In: *Mathematical Programming Computation* 1.1 (2009), pp. 1–41. DOI: 10.1007/s12532-008-0001-1.

[18] National Quantum Computing Centre. *2025 STFC Cross Cluster Proof of Concept: SparQ Quantum Computing Call*. Aug. 2025. URL: https://www.nqcc.ac.uk/funding/proof-of-concept-call-2025/.

[19] Gurobi Optimization LLC. *Gurobi Optimizer Reference Manual*. URL: https://docs.gurobi.com/projects/optimizer/en/current/index.html.

[20] Toshiki Teramura, Hiromi Ishii, and Taro Shimizu. *Jij-Inc/ommx: OMMX Python SDK 2.3.0*. Zenodo. Version python-2.3.0. DOI: 10.5281/zenodo.17638551. Nov. 2025. URL: https://doi.org/10.5281/zenodo.17638551.