Article

Scaling and networking a modular photonic quantum computer

https://doi.org/10.1038/s41586-024-08406-9

Received: 25 June 2024

Accepted: 14 November 2024

Published online: 22 January 2025

Open access

Check for updates

H. Aghaee Rad[1], T. Ainsworth[1], R. N. Alexander[1]✉, B. Altieri[1], M. F. Askarani[1], R. Baby[1], L. Banchi[1], B. Q. Baragiola[1], J. E. Bourassa[1], R. S. Chadwick[1], I. Charania[1], H. Chen[1], M. J. Collins[1], P. Contu[1], N. D'Arcy[1], G. Dauphinais[1], R. De Prins[1], D. Deschenes[1], I. Di Luch[1], S. Duque[1], P. Edke[1], S. E. Fayer[1], S. Ferracin[1], H. Ferretti[1], J. Gefaell[1], S. Glancy[1], C. González-Arciniegas[1], T. Grainge[1], Z. Han[1], J. Hastrup[1], L. G. Helt[1], T. Hillmann[1], J. Hundal[1], S. Izumi[1], T. Jaeken[1], M. Jonas[1], S. Kocsis[1], I. Krasnokutska[1], M. V. Larsen[1], P. Laskowski[1], F. Laudenbach[1], J. Lavoie[1]✉, M. Li[1], E. Lomonte[1], C. E. Lopetegui[1], B. Luey[1], A. P. Lund[1], C. Ma[1], L. S. Madsen[1], D. H. Mahler[1], L. Mantilla Calderón[1], M. Menotti[1], F. M. Miatto[1], B. Morrison[1], P. J. Nadkarni[1], T. Nakamura[1], L. Neuhaus[1], Z. Niu[1], R. Noro[1], K. Papirov[1], A. Pesah[1], D. S. Phillips[1], W. N. Plick[1], T. Rogalsky[1], F. Rortais[1], J. Sabines-Chesterking[1], S. Safavi-Bayat[1], E. Sazhaev[1], M. Seymour[1], K. Rezaei Shad[1], M. Silverman[1], S. A. Srinivasan[1], M. Stephan[1], Q. Y. Tang[1], J. F. Tasker[1], Y. S. Teo[1], R. B. Then[1], J. E. Tremblay[1], I. Tzitrin[1], V. D. Vaidya[1], M. Vasmer[1], Z. Vernon[1], L. F. S. S. M. Villalobos[1], B. W. Walshe[1], R. Weil[1], X. Xin[1], X. Yan[1], Y. Yao[1], M. Zamani Abnili[1] & Y. Zhang[1]

Photonics offers a promising platform for quantum computing[1–4], owing to the availability of chip integration for mass-manufacturable modules, fibre optics for networking and room-temperature operation of most components. However, experimental demonstrations are needed of complete integrated systems comprising all basic functionalities for universal and fault-tolerant operation[5]. Here we construct a (sub-performant) scale model of a quantum computer using 35 photonic chips to demonstrate its functionality and feasibility. This combines all the primitive components as discrete, scalable rack-deployed modules networked over fibre-optic interconnects, including 84 squeezers[6] and 36 photon-number-resolving detectors furnishing 12 physical qubit modes at each clock cycle. We use this machine, which we name Aurora, to synthesize a cluster state[7] entangled across separate chips with 86.4 billion modes, and demonstrate its capability of implementing the foliated distance-2 repetition code with real-time decoding. The key building blocks needed for universality and fault tolerance are demonstrated: heralded synthesis of single-temporal-mode non-Gaussian resource states, real-time multiplexing actuated on photon-number-resolving detection, spatiotemporal cluster-state formation with fibre buffers, and adaptive measurements implemented using chip-integrated homodyne detectors with real-time single-clock-cycle feedforward. We also present a detailed analysis of our architecture's tolerances for optical loss, which is the dominant and most challenging hurdle to crossing the fault-tolerant threshold. This work lays out the path to cross the fault-tolerant threshold and scale photonic quantum computers to the point of addressing useful applications.

Over the past 5 years, there has been a sea change in the focus of quantum-computing development efforts. Although the hardware available across all platforms is still firmly rooted in the noisy intermediate-scale quantum era[8], far in both performance and scale from the point of accessing high-value applications such as factoring[9] and quantum simulation of materials[10] or chemistry[11], there has been waning interest in exploiting such noisy intermediate-scale quantum machines to extract utility in the near term. Instead, research has turned towards advancing the state of the hardware supporting error correction and fault tolerance[1–4]. With recent resource estimates for promising algorithms requiring millions of gates applied to hundreds of logical qubits[12,13], a tactical retreat is warranted from near-term application implementation, to enable deeper investment of resources towards advancing scalability and physical qubit-level error rates.

The challenge of physically realizing a quantum computer that can deliver meaningful results on a useful algorithm hinges on two closely related hurdles: achieving component performance sufficient to yield physical qubit error rates that are below the threshold for fault tolerance[14–16], and the ability to scale the system to large numbers of qubits. Scaling is crucial not only to provide sufficient qubits to

[1]Xanadu Quantum Technologies Inc., Toronto, Ontario, Canada. ✉e-mail: rafael@xanadu.ai; jonathan@xanadu.ai

meet the demands of useful algorithms but also to accommodate the physical-to-logical qubit overhead (that is, encoding rate) required to suppress logical error rates to levels that are tolerable to the algorithm in question. The latter could be mitigated by using higher-rate quantum low-density parity-check (LDPC) codes[17], but remains a significant challenge. So far, none of the multiple strategies based on different physical substrates have overcome these hurdles, despite significant progress across many qubit modalities. Superconducting qubits have yielded demonstrations of computational advantage in random sampling problems[18,19], and it has been experimentally shown that error-correcting codes can be implemented in these machines to suppress error rates by increasing the code distance[3]. Neutral atom- and ion-trap-based platforms[4,20,21] have demonstrated logical gate implementation with convincing evidence of subthreshold operation. Within photonics-based platforms, machines have hosted sampling-based demonstrations of quantum computational advantage[22,23], although they suffer from high photon losses and other noise sources that make them vulnerable to classical simulation[24], as well as a wide array of programmable quantum information processing tasks[25,26]. Here we design and demonstrate a complete photonic architecture that can, once appropriate component performance is achieved, deliver a universal and fault-tolerant quantum computer.

To achieve quantum computing, photonics platforms require the development of an architecture that comprehensively addresses all aspects of qubit synthesis, control and measurement in the context of fault-tolerant operation. Even notwithstanding performance, the existing demonstrations of photonic quantum computers[22,23,25], although groundbreaking in their own right, all lack key functional features that are required to furnish a universal machine capable of implementing qubit error correction and fault-tolerant gates. Implementations of single-photon-based dual-rail-encoded qubit architectures[26] have so far failed to incorporate the multiplexing subsystems needed to overcome punishingly low success probabilities in qubit synthesis and non-deterministic gates, and lack the features necessary for real-time diagnosis and correction of error syndromes. Although the performance of individual photonic components is still too limited by optical loss to operate in the fault-tolerant regime, demonstrations of the functionality of these components and the platform and systems integration needed to scale them need not wait.

Alongside progress in component performance, it is critical to characterize the evolving requirements for achieving fault tolerance and translate these into a detailed mapping between the high-level functions of the architecture and the physical building blocks used to implement them. This enables optimization of configurations with respect to a performance model constrained by realistic hardware limitations, accelerating progress. In the process, it is essential to include advances in quantum error correction that relax the requirements for fault tolerance—here we incorporate and report on decoder-based improvements to the quantum-error-correction threshold. Earlier examples of photonic architectures proposed laid out the abstract basic building blocks needed, namely, sources of few-photon resource states and spatiotemporal linear optical operations augmented by single-photon detectors for the 'fusion-based' approach[27], or sources of non-Gaussian states and spatiotemporal linear optical operations augmented by homodyne detectors for the optical Gottesman–Kitaev–Preskill (GKP) approach[5,28–31]. Although promising progress on the performance and function of many building blocks for both approaches has been reported[6,25,26,32–34], no complete photonic architecture has been experimentally demonstrated in practice at any scale, leaving claims of modularity, networkability and scalability open to speculation.

The architecture presented here follows the optical GKP approach, which offers a distinct advantage in its ability to implement logic gates and error correction using deterministic, room-temperature linear optical operations and modest component depths for the various optical paths in the system. Entangling operations and logic gates are deterministic, relying on beamsplitters and photodiodes (that is, only room-temperature components) to furnish the physical functions necessary; this is to be contrasted with the single-photon approach, which suffers from non-deterministic operation and requires (cryogenic) superconducting photon detectors not only for input state synthesis but also at almost every stage. In comparison, the optical GKP approach requires cryogenics only to herald certain input states at the qubit preparation stage.

The original continuous-variable measurement-based quantum computing model[35] is similar to its qubit counterpart. Key features are that information is encoded in the quadrature basis and Gaussian unitaries and homodyne measurements have the role of Clifford gates and Pauli measurements. The model can be made fault tolerant while preserving these features by encoding qubits in each mode through the GKP code[36]. Further details about fault tolerance and universality can be found in ref. 5. By switching to resource states based on macronodes constructed from entangled pairs and Greenberger–Home–Zeilinger (GHZ)-type measurements, this model becomes easier to implement (only requires photon-number-preserving Gaussian unitaries) and has better performance[29,37].

Our architecture (including the Aurora system) is shown in Fig. 1, and consists of three stages. Gaussian boson sampling (GBS) devices prepare the heralded initial non-Gaussian states. Adaptive interferometer trees with homodyne detectors (which we refer to as 'refineries') improve the quality and probability of the non-Gaussian states and entangling them into two-mode GKP Bell pairs. This is followed by an array of quantum processing unit (QPU) cells that select the best-quality Bell pairs, entangle them into a spatiotemporal cluster state and implement gates by performing homodyne measurements on each mode; here we use the term 'QPU' to refer strictly to this subsystem and its constituent cells, not the entire quantum computer apparatus. Each of these three stages is implemented on distinct sets of photonic integrated circuit (PIC) chips, which are networked with phase- and polarization-stabilized fibre-optical interconnects.

Using GBS devices unassisted to directly produce GKP states has the drawback of low success probability[38–40], introducing excessive levels of loss and physical component overheads by requiring high-depth switch networks. This can be mitigated by searching the GBS circuit parameter space and using configurations found that furnish useful non-Gaussian states with higher probability. These states need not be GKP states themselves, provided they can be converted into GKP states at the refinery. In our architecture, as depicted in Fig. 1, the refinery contains two symmetric binary trees of adaptive beamsplitters, with all but one of the outputs of each tree being measured in the momentum quadrature basis by homodyne detectors.

Each unit cell of the binary tree can perform either a simple switch operation or one round of breeding[41], enabling any subset of the input states to be selected and bred, as the breeding steps can occur anywhere in the tree. This is equivalent to using an N-inputs-to-M-outputs multiplexer (MUX), followed by breeding of the M outputs, but with a shallower optical path and thus less loss. The outputs of the trees are GKP states defined on rectangular lattices, specified by the photon-number outcomes. A measurement-based squeezer[42], using only elements already required (squeezed states provided on demand by select outputs of the GBS chips, an adaptive beamsplitter and a homodyne detector), is then used on each tree output to align them to a single GKP phase space lattice. The chosen lattice corresponds to the so-called square-grid qunaught state, which is $\sqrt{2\pi}$ periodic in both position and momentum quadratures[43,44]. Finally, the outputs of two such trees are interfered on a 50:50 beamsplitter that entangles them to form a GKP Bell pair, which serves as the basic unit from which a cluster state can be synthesized.

The choice of cluster state and error-correction code used simply correspond to a choice of how the fibres carrying the Bell pairs are routed to the next array of QPU chips. Thus, implementing lattices

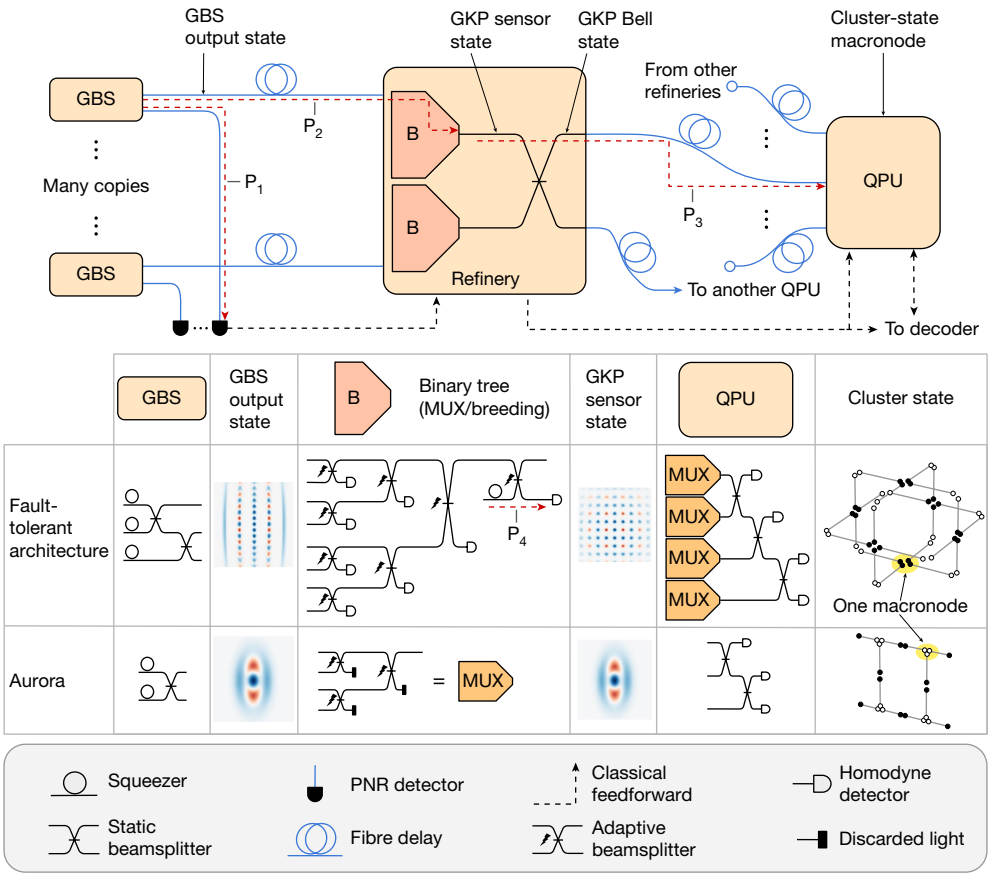


Fig. 1 | Layout of the architecture including loss paths P_1, P_2, P_3 and P_4. Top: schematic of our architecture. Precursors to GKP states are generated from multimode Gaussian states produced probabilistically with GBS chips by heralding particular PNR patterns. Many precursor states are sent to each refinery chip (via optical fibre delays represented by blue lines), which use a combination of multiplexing and breeding implemented in a binary tree of beamsplitters (represented by the wedge-like shapes labelled 'B'), and squeezing, to create a pair of high-quality GKP sensor states. For the fault-tolerant architecture, the binary tree is augmented with homodyne detectors. A Bell pair is then generated by applying a 50/50 beamsplitter (black solid lines). The spatial routing and temporal delays of the modes in each pair are set by the desired cluster-state graph, such that each graph macronode corresponds to an individual QPU chip and these chips share entangled pairs if they are neighbours on the cluster-state graph. In the fault-tolerant architecture, multiple pairs are created per edge, but only one pair is selected per graph edge by a multiplexer at the beginning of the QPU. Then, each QPU interferes with the selected pairs using static beamsplitters and these modes are measured using homodyne detection. Loss paths are shown with red dashed lines while classical feedforward is represented with black dashed lines. Middle: table showing the internal structure of each submodule in the fault-tolerant architecture and in the Aurora experiment. Bottom: legend for optical component diagrams.

with non-local connectivity is straightforward, making our architecture compatible with higher-rate LDPC codes[17]. To ensure persistent entanglement between qubit sites at adjacent clock periods—required for measurement-based quantum computation—a subset of Bell pairs experiences a time delay on one of their modes using a fibre-optical delay line. Several GKP Bell pairs are generated per cluster-state lattice edge, with the two modes from each pair sent to QPU cells. Each of the QPU cells, which are arranged in arrays on another set of photonic chips, corresponds to a cluster-state lattice site, or macronode[7]; these sites in turn correspond to physical qubits available for computation. The first stage of the QPU chips provides a final layer of switching, where the best pair per lattice edge is selected for use by a small binary tree; this selection is made based on the homodyne detection outcomes in the refinery, and the photon-counting outcomes from the GBS cells. The selected Bell pairs are finally subjected, within each QPU cell, to phase shifters that implement GKP Hadamard gates, converting them into two-qubit cluster states, and then a short sequence of static beamsplitters and homodyne measurements that project these inputs onto GHZ states[29,37] to create the desired fully connected cluster state[37]. The measurement bases are selected on each clock cycle by a classical controller that is informed by both the user-defined algorithm and the error-correction protocol (including a decoder), taking into account measurement outcomes from previous computational time steps. Full universality is achieved with magic states encoded in the GKP code and included into the pair creation[29] or generated through measurements on the cluster state[45]. Such magic-state generation is not expected to worsen the tolerance for losses in our architecture, but further work is needed to conclusively verify this.

Experiment

The architecture described relies on four key functionalities, all of which must be implemented in an integrated and indefinitely scalable platform to furnish a fault-tolerant quantum computer without inherent limitation in qubit number. These functionalities are: (1) heralded synthesis of non-Gaussian states using photon-number-resolving (PNR) detectors, (2) real-time feedforward actuation of binary trees of beamsplitters based on these detector events, (3) entanglement of the outputs of these trees to form a spatiotemporal cluster state, and (4) quadrature measurements implemented on all nodes of the cluster state at each clock period fed into a decoder implemented in real time, with single-clock-cycle feedforward available to inform subsequent measurement bases using prior measurement outcomes. To demonstrate the technological feasibility of this approach, we constructed a

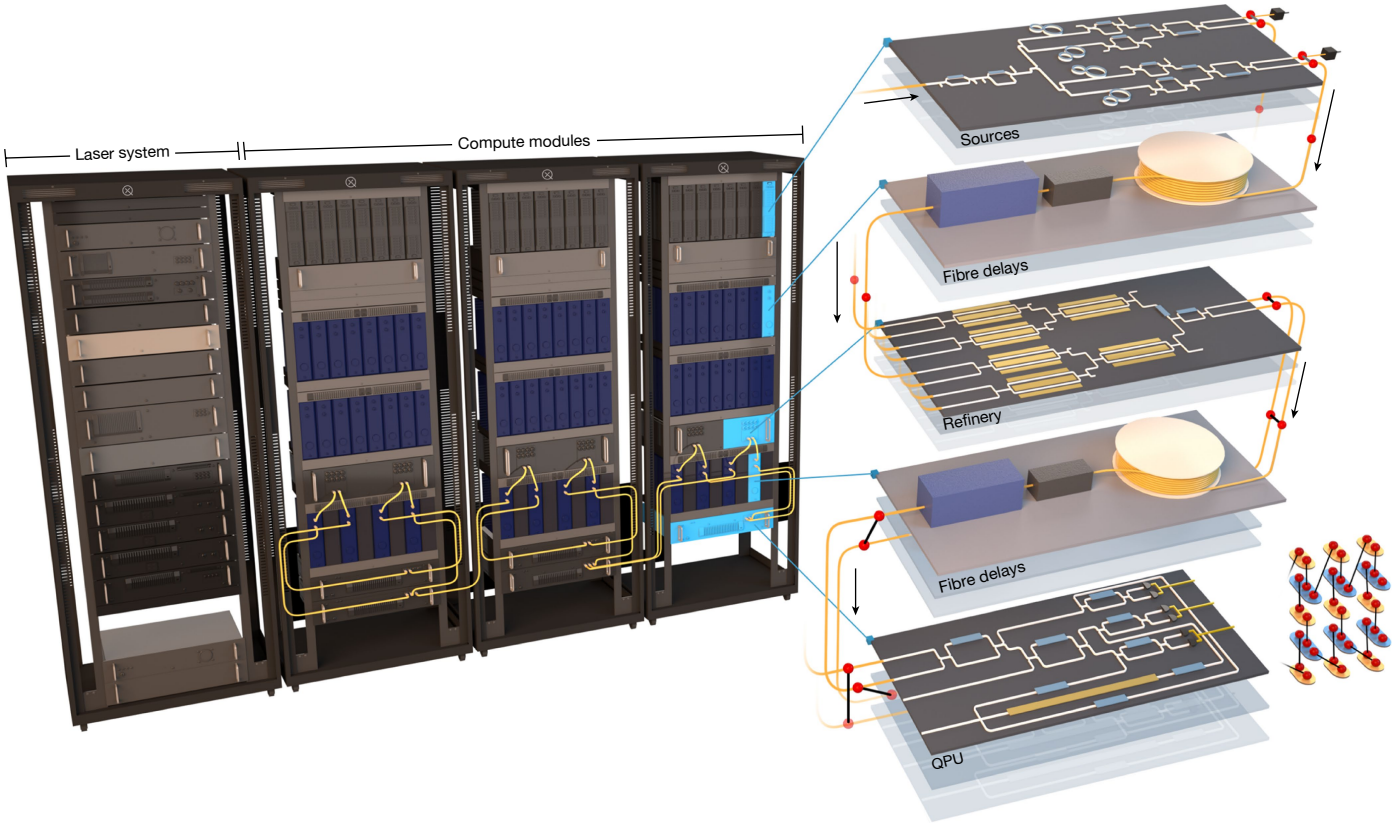


Fig. 2 | Schematic diagram of the Aurora system and main modules. An array of 24 sources chips generates squeezed states and entangled two-mode Gaussian states. These are pumped by a customized pulsed laser system (leftmost rack), which also generates and distributes local oscillator beams and reference beams for locking to the compute modules. PNR detectors are used on one half of each of the two-mode Gaussian state outputs from these chips (sources) to herald a non-Gaussian state; a stabilized fibre delay line buffers (fibre delays) the other mode while awaiting these detection results. The heralded outputs are fed into an array of six refinery chips (refinery), each of which is dynamically actuated to select the best-available pair of inputs using a pair of four-to-one binary tree multiplexers to synthesize an entangled Bell pair. Six such pairs are available after the refinery; one half of two pairs is delayed, through the routing modules (fibre delays), to generate entanglement between adjacent clock periods. All pairs are then stitched into a spatiotemporal cluster state by an array of 5 QPU chips, which also performs homodyne measurements on all 12 operating modes on every clock cycle. Equivalently, each QPU chip implements a multimode GHZ measurement, thereby generating a fully connected resource state. The fibre routing pattern between the refinery and the QPUs, illustrated by the yellow cabling at the bottom of the racks, implements the desired cluster-state lattice by appropriately networking the QPUs. For clarity, other fibre cabling paths, from the laser system to the compute system and from the sources modules to PNRs or down to the refinery inputs through delay lines, are not shown. Details of the full-system hardware blueprint, the laser system and modules can be found in Supplementary Figs. 22, 24 and 30–33, respectively.

model of this architecture (shown in Fig. 1, using the components listed in the 'Aurora' row of the table) incorporating all of these functional features in a modular, indefinitely scalable platform. This machine hosts a sources subsystem with 84 squeezers within 42 GBS cells, distributed across an array of 21 (plus 3 extras for redundancy) silicon-nitride PICs, providing 12 squeezed states and 36 heralded non-Gaussian states using 36 PNR detectors. The outputs feed 48 inputs to a refinery array with 12 binary switch trees across 6 thin-film lithium-niobate multiplexer PICs, each yielding 1 entangled Bell pair. The modes comprising these pairs, after suitable temporal delay on a subset, are entangled into a cluster state and measured on each clock cycle by five QPU chips (through the method described in ref. 37), implemented on silicon PICs, which interface in real time with a classical decoder implemented on a field-programmable gate array. A schematic of the machine is presented in Fig. 2 and is described in full detail in Supplementary Information. The entire system fits into four standard server racks that house fully packaged modules operating at room temperature, with the sole exception of the PNR detection system, which is housed in a cryostat. All PIC modules are networked using custom phase- and polarization-stabilized fibre delay line modules to carry quantum light between stages and appropriately entangle modes on separate QPU chips to enable cluster-state synthesis.

To benchmark the key functionalities 1–4 listed above, 2 core experiments were carried out using the machine. First, the system was programmed to pass only squeezed states through to the QPU array (as opposed to the heralded non-Gaussian state outputs) and thus synthesize a $12 \times N$-mode Gaussian cluster state, where N is the duration of the experiment in clock periods. This benchmark tests all functional components of the system except feedforward features and PNR detection. The description of the resource state—prepared using the methods described in ref. 37—can be simplified by back-evolving the measured modes through the QPU unitary (also known as expressing the state in terms of its distributed modes[44]), resulting in correlated pairs arranged as shown in Fig. 3a. Neighbouring macronodes are connected if the pair that links them is inseparable. The results (nullifier variances) are shown in Fig. 3b, plotted as a function of their temporal mode index. Continuously acquired over 2 hours, this result represents the synthesis and measurement of a macronode cluster state consisting of 86.4 billion modes, or 86.4/12 = 7.2 billion temporal modes. Despite the high optical losses of about 14 dB from the synthesis of the squeezed states to their ultimate detection in the QPU, these variances are persistently below the vacuum noise level, indicating squeezing and validating the entanglement present in the cluster between both modes that form each pair. The degree of

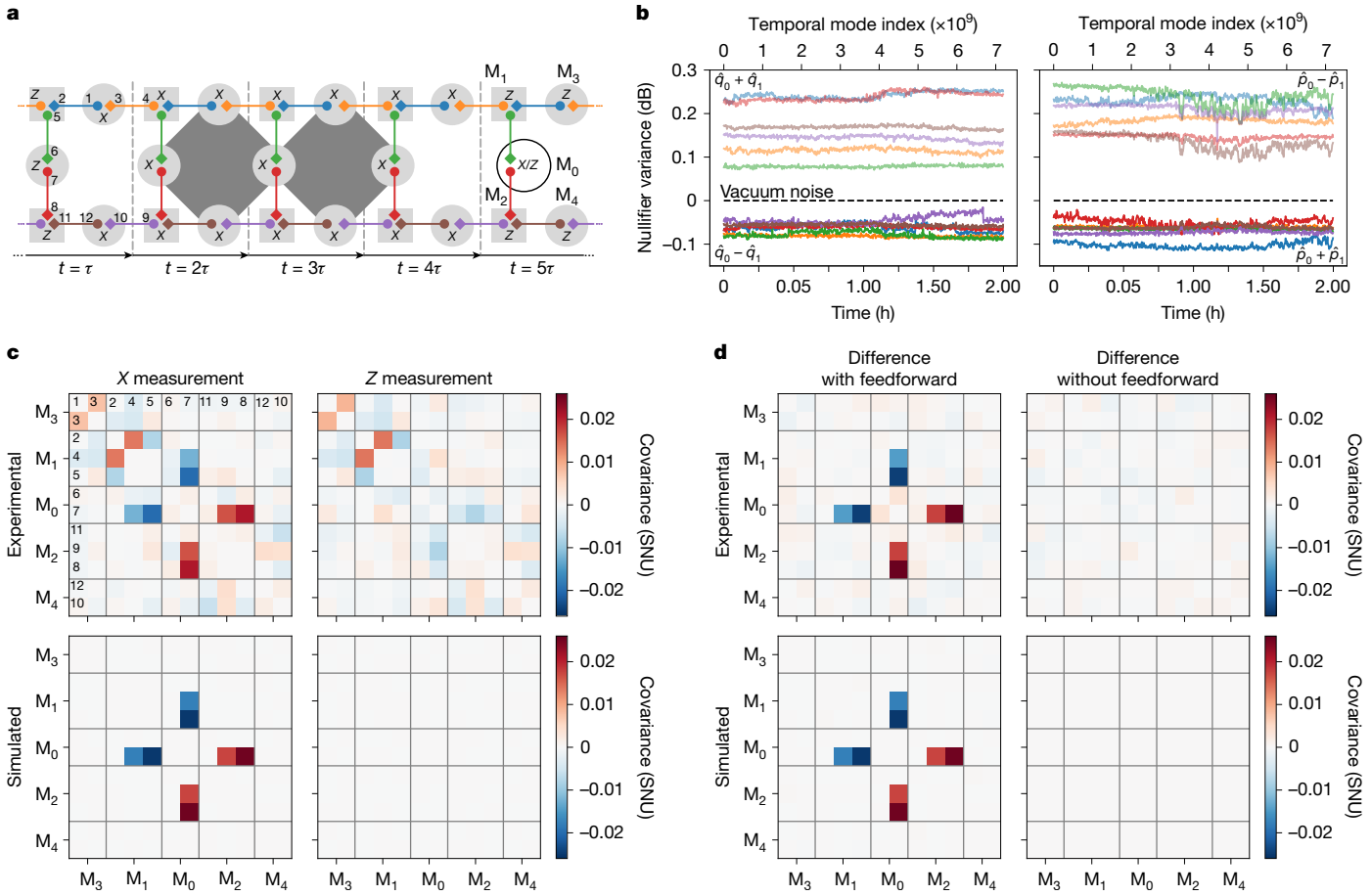


Fig. 3 | Synthesis and measurement of a macronode cluster state. a, Graph representation of the 12-mode cluster state. For each time window (t), entangled pairs (coloured dots joined by a solid line) span neighbouring macronodes. Mode (macronode) labels are provided in the first (fifth) time step and τ is the experiment clock period. One homodyne measurement per macronode is programmable, the rest are measured in $\hat{q}$. For mode pairs (3, 4) and (9, 10), one mode has been time delayed, allowing computation and entanglement to persist in time. $\hat{q}$ and $\hat{p}$ represent the position and momentum quadrature operator, respectively. **b**, Nullifier variances versus time. By applying the inverse linear transformation of the macronode beamsplitter network to the measured quadrature outcomes, we obtained the quadrature values corresponding to the six individual entangled pairs before mutual interference. The nullifiers of these states (solid lines) are of the form $\hat{q}_0 - \hat{q}_1$ (left panel) and $\hat{p}_0 + \hat{p}_1$ (right panel). Their variance is kept below the vacuum noise over a continuous acquisition of 2 h (corresponding to 86.4 billion modes, the product of 7.2 billion temporal modes over 12 spatial modes) at a clock rate of 1 MHz. The operators $\hat{q}_0 + \hat{q}_1$ (left panel) and $\hat{p}_0 - \hat{p}_1$ (right panel) are operators relating to the anti-squeezed directions of the two-mode squeezed state, and their variances are observed (faint lines) to be above vacuum noise for the duration of the acquisition. **c**, Adaptive capability. The X, Z and X/Z labels in **a** specify a measurement basis for the programmable homodyne measurement in each macronode, corresponding to $\hat{p}$, $\hat{q}$ or either, respectively. This measurement pattern measures two four-body repetition code check operators (with support on the two dark grey diamonds in **a**). This measurement pattern is repeated and homodyne data are collected for the 12 modes in every fifth time step. Separate covariance matrices are constructed from data when mode M_0 is measured in X (left) versus Z (right). Matrix rows and columns are labelled to reflect the numbering of modes in **a**. The single-mode variance on the diagonals was removed for better contrast, while the covariance values are plotted in shot-noise units (SNU), the units in which vacuum has a variance of 1 and $\hbar = 2$. **d**, Control experiment. The difference between covariance matrices corresponding to the X- and Z-measurement basis decoder decision shown in **c** (left, 'with feedforward'), and from an otherwise identical experiment in which the basis measured is determined by random choice instead of from the decoder decision (right, 'without feedforward').

nullifier squeezing agrees well with numerical simulations of the machine.

To showcase the feedforward and non-Gaussian-state synthesis capabilities of our device, a repetition code error-detection experiment on low-quality GKP states was carried out. Heralded on detecting two photons in the GBS heralding mode, outputs of GBS cells are found in even-parity squeezed cat states that crudely approximate simple two-peaked GKP sensor states[43]. On failure to observe two photons, a squeezed state is selected by the multiplexer, so that the resulting cluster-state modes are in a combination of squeezed states and squeezed cat states—the latter make up 3.05% on average of the modes. These are sufficiently rare that our results do not depend on their inclusion; the results are nearly identical if only squeezed states are used. We initialize a computation that performs two (foliated) repetition code checks (through the measurements in time steps 1–4 in Fig. 3a). Outcomes are processed by the QPU decoder, which computes bit values and estimates of phase error probabilities associated with the qubits involved in the parity checks, which are in turn used in belief propagation decoding[46]. The decoder outputs an updated distribution of error probabilities used to quantify confidence in the recovery. We apply a thresholding function to the confidence value to decide which basis to measure at the following time step—if the recovery was established with high (low) confidence, the macronode M_0 in time step 5 is measured in the X (Z) basis. The thresholding function results in X and Z being selected 51.206% and 48.794% of the time, respectively. The effect of the feedforward operation can be observed by comparing correlations present in the homodyne data from modes in the fifth time step between X and Z selected cases (Fig. 3c). We measure Z on macronodes

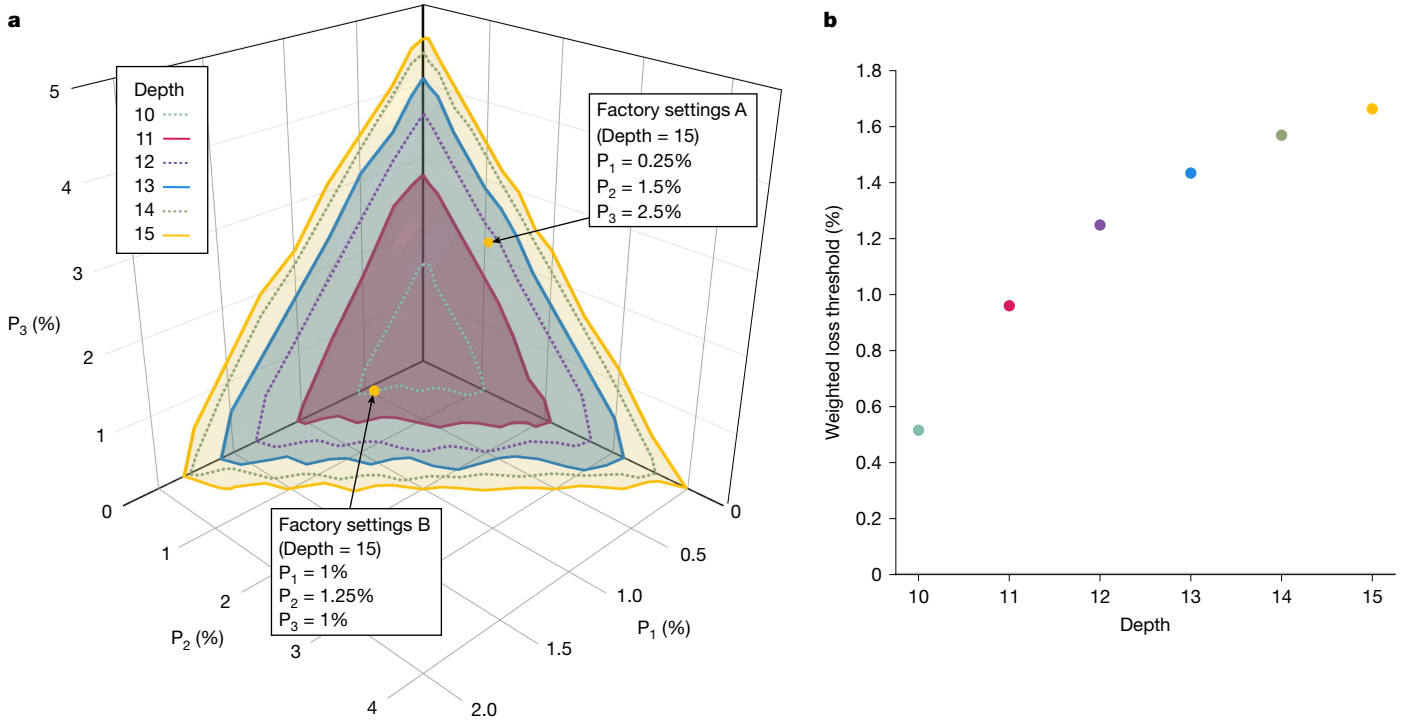


Fig. 4 | Fault-tolerance threshold. a, Loss budget and multiplexing depth required for fault-tolerant operation. We analyse loss tolerance along the three paths P_1, P_2 and P_3 as a function of the combined depth of balanced trees of beamsplitters in refinery and QPU chips shown in Fig. 1. For architecture configurations with a given combined depth, surfaces are plotted that have been fitted to the most extremal attainable P_1, P_2 and P_3 tolerance values. Thus, each point on a given surface represents a combination of loss budgets assuming fixed total depth for the three principal optical paths in the system to meet the requirements for fault tolerance. The two yellow dots correspond to a pair of example configurations with combined refinery and QPU depths of 15. **b**, Weighted loss threshold and depth. Changing the depth affects the distance of the surfaces in **a** from the origin. For each depth, we plot the weighted loss threshold, defined to be the average length component of all surface data points in the direction of an approximate normal vector $\mathbf{v} = (0.85, 0.41, 0.34)$. The vector $\mathbf{v}$ is found by first constructing normal vectors for the planes specified by the three extremal 'corner' points of each surface, and then taking the average. Analysis of the architecture configurations with two-mode GBSs is presented in the Supplementary Information.

M_1–M_5 and restrict two of the dumbbells, modes (7, 8) and modes (5, 6) in Fig. 3a, to always be Gaussian states so that the quadrature data have a simpler and more distinct structure. If the measurement readout and decoding could not be completed in time for the fifth step in the protocol, then the basis would not be chosen correctly with odds better than approximately 50/50 random. We can simulate this in our set-up by repeating the protocol with the M_0 basis in the fifth time step chosen by a binary random variable uncorrelated with the decoding decision. If homodyne correlations were again plotted based on the decoding decision, this time they should appear identical. We plot the difference between the decoder-outcome-determined covariance matrices for the feedforward and no-feedforward scenarios in Fig. 3d.

Loss requirements

Returning to the full architecture proposed, we note that fault tolerance in any photonic architecture is highly dependent on photon loss, which is present in all optical submodules and components in Fig. 1. It is noted that sequential loss channels can be combined (transmissivity decays multiplicatively) and uniform loss along multiple paths commutes with linear optical transformations. Thus, the loss contributions can be converted into a single-mode channel acting on all modes independently, and this can be commuted to act before photon-counting or homodyne detectors, or any other chosen point in the path. We define three primary optical paths: P_1, from squeezer to photon counter, P_2, from squeezer to the homodyne in the refinery, and P_3 from the output of the MUX and breeding tree in the refinery to immediately before the homodyne detectors in the QPU. This way, the loss of the longest path can be computed by combining P_2 and P_3 losses (assuming that the detector losses in the refinery are equal to those in the QPU). We also define the minor loss path P_4, for the mode used in measurement-based squeezing. These loss paths are shown in Fig. 1. Commuting the relevant losses appropriately, their effects can be captured fully by figures of merit associated with the GKP Bell pair units that comprise the basic units of the architecture. The quality of approximate GKP qubits can be quantified by how well they act as physical qubits for the next layer of quantum error correction. We do this by relating the symmetric effective squeezing (see Supplementary Information for a description) to the squeezing requirements for fault tolerance of the surface code concatenated with the GKP code[47].

We perform a comprehensive optimization of different candidate 'state factories'—the collections of elements that provide the best-available GKP Bell pairs to be entangled into a cluster state—and present the results in Fig. 4, in which the fault-tolerance threshold is visualized in relation to the loss present in the three principal optical paths in the architecture. The details of this optimization are presented in Methods. For ease of visualization, P_4 is not represented. However, P_4 incorporates a small subset of the elements experienced in P_3, and its loss requirements are less stringent than any of P_1–P_3; thus if the loss bounds shown for those paths are achievable, so too must be the loss bound for P_4. Each plane represents a fixed total depth (refinery tree depth plus QPU tree depth), and separates the region (in loss parameter space) compatible with fault-tolerant operation, closer to the origin, from that of above threshold operation. This puts clear lower bounds, for the architecture considered, on how much loss can be tolerated in the components comprising each optical path, given a constraint (dictated by engineering and cost requirements) placed on the number of GBS cells and maximum allowable refinery chip inputs

per physical qubit. In turn, this provides benchmarks against which hardware choices—such as around waveguide materials, chip–fibre coupling schemes and component designs—can be qualified.

Discussion and outlook

Although demonstrations of systems like Aurora help build confidence in the scalability of the photonic approach to quantum computing, there remains a gap—as with all hardware approaches—between present-day performance and the demands of fault tolerance. Although refinements are necessary to carry the modules responsible for qubit synthesis and processing from prototypes to mass manufacturing, our results indicate that the present-day technological backdrop of photonic-chip fabrication, classical control electronics and fibre-optical networking make feasible the task of modularizing and scaling a realistic photonic architecture for fault-tolerant quantum computing. In addition, the quantum optical theoretical underpinnings are now sufficiently well developed to enable thorough optimization of optical GKP-based architectures to find those that are most hardware efficient and tolerant to physical imperfections.

The component performance gap, however, is significant: whereas the currently known optimal configurations demand about 1% loss budgets and challenging about 10 multiplexing depths to achieve fault-tolerant operation, Aurora shows losses of about 56% for the heralding paths (P_1), and slightly over 95% for the heralded optical paths (P_1 and P_2). Indeed, this system was constructed to demonstrate the scalability of the approach through modularity and networking. Although extensive design iterations and post-fabrication selection were performed, no special loss optimizations were carried out on the chip platforms employed, which were all based on pre-existing commercially available fabrication lines.

The loss tolerances in Fig. 4 are not hard upper bounds but rather are attainable lower bounds by explicit circuits; here we limited ourselves to a structured analytically derived family of GBS sources, adaptive cat-state breeding as the refinery protocol and a particular method for constructing cluster states. The loss tolerance of paths P_1, P_2, and P_3 can be increased and circuit depths of refinery and QPU chips can be decreased by further optimization of theoretical protocols, although we expect these loss tolerances to remain within one order of magnitude of currently contemplated ranges. This may be accomplished, for example, by using alternative GBS circuits, more advanced refinery protocols, alternative cluster-state-generation protocols, noise-tailored quantum-error-correcting codes along with noise-biased or non-square lattice GKP codes[37], and closer-to-optimal decoders, all of which are deserving of future attention and may modify the loss tolerances of different paths by multiple percentage points.

Closing the gap between the present state of the art in hardware components and that required for fault tolerance is expected to require contribution from both architectural refinements and hardware improvements. Towards the latter, intensive efforts are underway in customized fabrication process engineering and photonic and fibre component design to achieve the loss budgets required. We summarize our latest component-level results towards this goal for each relevant subsystem in Methods. Taken in aggregate, an improvement of 20–30 times (when measured on a decibel scale) in each photonic component insertion loss compared with this current state of the art would enable fault-tolerant operation even if no further progress is made in relaxing architectural requirements. Recent reports of platform development for comparable components[26] have also shown promising progress in loss reduction.

Beyond merely achieving component performance compatible with fault tolerance, manufacturing methods must be developed to ensure that these performance levels can be maintained in the context of mass production. For example, a depth-12 machine (that is, one with 8-times improved state factory overhead versus that considered in Fig. 4) with 100 logical qubits and a 100-to-1 error-correction overhead would require tens of millions of GBS cells. Even assuming a 100-times improvement in state-factory component density, this will necessitate tens of thousands of server racks in the state factory; if not all of these assumed improvements are possible, even more may be required. Although this is well within the scale of present-day classical data centres, progress in quantum photonic component performance must be guided by constraints that respect these requirements for large-scale manufacturing.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-08406-9.

1. Bravyi, S. et al. High-threshold and low-overhead fault-tolerant quantum memory. *Nature* **627**, 778–782 (2024).
2. Paetznick, A. et al. Demonstration of logical qubits and repeated error correction with better-than-physical error rates. Preprint at https://arxiv.org/abs/2404.02280 (2024).
3. Google Quantum AI. Suppressing quantum errors by scaling a surface code logical qubit. *Nature* **614**, 676–681 (2023).
4. Bluvstein, D. et al. Logical quantum processor based on reconfigurable atom arrays. *Nature* **626**, 58–65 (2024).
5. Bourassa, J. E. et al. Blueprint for a scalable photonic fault-tolerant quantum computer. *Quantum* **5**, 392 (2021).
6. Zhang, Y. et al. Squeezed light from a nanophotonic molecule. *Nat. Commun.* **12**, 2233 (2021).
7. Menicucci, N. C. Temporal-mode continuous-variable cluster states using linear optics. *Phys. Rev. A* **83**, 062314 (2011).
8. Preskill, J. Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
9. Shor, P. W. Algorithms for quantum computation: discrete logarithms and factoring. In *Proc. 35th Annual Symposium On Foundations of Computer Science* 124–134 (IEEE, 1994).
10. Alexeev, Y. et al. Quantum-centric supercomputing for materials science: a perspective on challenges and future directions. *Future Gener. Comput. Syst.* **160**, 666–710 (2024).
11. McArdle, S., Endo, S., Aspuru-Guzik, A., Benjamin, S. C. & Yuan, X. Quantum computational chemistry. *Rev. Mod. Phys* **92**, 015003 (2020).
12. Gidney, C. & Ekerå, M. How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits. *Quantum* **5**, 433 (2021).
13. Dalzell, A. M. et al. Quantum algorithms: a survey of applications and end-to-end complexities. Preprint at https://arxiv.org/abs/2310.03011 (2023).
14. Aharonov, D. & Ben-Or, M. Fault-tolerant quantum computation with constant error. In *Proc. 29th Annual ACM Symposium on Theory of Computing* 176–188 (ACM, 1997).
15. Kitaev, A. Y. Quantum computations: algorithms and error correction. *Russ. Math. Surv.* **52**, 1191 (1997).
16. Knill, E., Laflamme, R. & Zurek, W. H. Resilient quantum computation. *Science* **279**, 342–345 (1998).
17. Breuckmann, N. P. & Eberhardt, J. N. Quantum low-density parity-check codes. *PRX Quantum* **2**, 040101 (2021).
18. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).
19. Wu, Y. et al. Strong quantum computational advantage using a superconducting quantum processor. *Phys. Rev. Lett.* **127**, 180501 (2021).
20. Moses, S. A. et al. A race-track trapped-ion quantum processor. *Phys. Rev. X* **13**, 041052 (2023).
21. Egan, L. et al. Fault-tolerant control of an error-corrected qubit. *Nature* **598**, 281–286 (2021).
22. Madsen, L. S. et al. Quantum computational advantage with a programmable photonic processor. *Nature* **606**, 75–81 (2022).
23. Zhong, H.-S. et al. Quantum computational advantage using photons. *Science* **370**, 1460–1463 (2020).
24. Oh, C., Liu, M., Alexeev, Y., Fefferman, B. & Jiang, L. Classical algorithm for simulating experimental Gaussian boson sampling. *Nat. Phys.* **20**, 1461–1468 (2024).
25. Arrazola, J. M. et al. Quantum circuits with many photons on a programmable nanophotonic chip. *Nature* **591**, 54–60 (2021).
26. Alexander, K. et al. A manufacturable platform for photonic quantum computing. Preprint at https://arxiv.org/abs/2404.17570 (2024).
27. Bartolucci, S. et al. Fusion-based quantum computation. *Nat. Commun.* **14**, 912 (2023).
28. Gottesman, D., Kitaev, A. & Preskill, J. Encoding a qubit in an oscillator. *Phys. Rev. A* **64**, 012310 (2001).
29. Tzitrin, I. et al. Fault-tolerant quantum computation with static linear optics. *PRX Quantum* https://doi.org/10.1103/PRXQuantum.2.040353 (2021).
30. Larsen, M. V., Chamberland, C., Noh, K., Neergaard-Nielsen, J. S. & Andersen, U. L. Fault-tolerant continuous-variable measurement-based quantum computation architecture. *PRX Quantum* **2**, 030325 (2021).
31. Fukui, K., Asavanant, W. & Furusawa, A. Temporal-mode continuous-variable three-dimensional cluster state for topologically protected measurement-based quantum computation. *Phys. Rev. A* **102**, 032614 (2020).

32. Asavanant, W. et al. Time-domain-multiplexed measurement-based quantum operations with 25-MHz clock frequency. *Phys. Rev. Appl.* **16**, 034005 (2021).
33. Larsen, M. V., Guo, X., Breum, C. R., Neergaard-Nielsen, J. S. & Andersen, U. L. Deterministic multi-mode gates on a scalable photonic quantum computing platform. *Nat. Phys.* **17**, 1018–1023 (2021).
34. Konno, S. et al. Logical states for fault-tolerant quantum computation with propagating light. *Science* **383**, 289–293 (2024).
35. Menicucci, N. C. et al. Universal quantum computation with continuous-variable cluster states. *Phys. Rev. Lett.* **97**, 110501 (2006).
36. Menicucci, N. C. Fault-tolerant measurement-based quantum computing with continuous-variable cluster states. *Phys. Rev. Lett.* **112**, 120504 (2014).
37. Walshe, B. W. et al. Linear-optical quantum computation with arbitrary error-correcting codes. Preprint at https://arxiv.org/abs/2408.04126 (2024).
38. Tzitrin, I., Bourassa, J. E., Menicucci, N. C. & Sabapathy, K. K. Progress towards practical qubit computation using approximate Gottesman–Kitaev–Preskill codes. *Phys. Rev. A* https://doi.org/10.1103/PhysRevA.101.032315 (2020).
39. Fukui, K. et al. Efficient backcasting search for optical quantum state synthesis. *Phys. Rev. Lett.* **128**, 240503 (2022).
40. Takase, K. et al. Gottesman–Kitaev–Preskill qubit synthesizer for propagating light. *npj Quantum Inf.* https://doi.org/10.1038/s41534-023-00772-y (2023).
41. Weigand, D. J. & Terhal, B. M. Generating grid states from Schrödinger-cat states without postselection. *Phys. Rev. A* **97**, 022341 (2018).
42. Filip, R., Marek, P. & Andersen, U. L. Measurement-induced continuous-variable quantum interactions. *Phys. Rev. A* **71**, 042308 (2005).
43. Duivenvoorden, K., Terhal, B. M. & Weigand, D. Single-mode displacement sensor. *Phys. Rev. A* **95**, 012305 (2017).
44. Walshe, B. W., Baragiola, B. Q., Alexander, R. N. & Menicucci, N. C. Continuous-variable gate teleportation and bosonic-code error correction. *Phys. Rev. A* **102**, 062411 (2020).
45. Walshe, B. W., Alexander, R. N., Menicucci, N. C. & Baragiola, B. Q. Streamlined quantum computing with macronode cluster states. *Phys. Rev. A* **104**, 062427 (2021).
46. Richardson, T. & Urbanke, R. *Modern Coding Theory* (Cambridge Univ. Press, 2008).
47. Marek, P. Ground state nature and nonlinear squeezing of Gottesman–Kitaev–Preskill states. *Phys. Rev. Lett.* **132**, 210601 (2024).

Methods

The Aurora hardware system (Extended Data Fig. 1) is composed of six distinct principal subsystems: (1) a customized master laser system provides coherent pump and local oscillator beams, as well as reference beams for phase stabilization; (2) a sources array generates squeezed light and two-mode Gaussian states; (3) a PNR detection system is used for heralding non-Gaussian states; (4) an array of refineries, each multiplexing eight inputs to one entangled pair; (5) a QPU array forms the spatial and temporal connections in the cluster state and performs homodyne measurements on each qubit; and (6) an array of fibre buffers provides appropriate phase- and polarization-stabilized delay lines between the sources and refineries, as well as between the refineries and QPUs. The entire system, apart from the cryogenic detection array, fits into 4 standard 19-inch server racks. Here we summarize the main features of these subsystems, as well as our method for verifying the multimode entanglement present in our cluster-state benchmark experiment; a more detailed exposition of these topics is available in Supplementary Information.

Laser system

The laser system is responsible for providing appropriate pump pulses (P1 and P2) to each squeezer in the sources array, a local oscillator beam temporally mode-matched to the quantum pulses for homodyne detection, and a variety of reference beams used to stabilize fibre delays (ref) and resonator positions (probe) throughout the rest of the system.

The laser system (depicted in Supplementary Fig. 24) begins with five narrow-linewidth lasers (P1, P2, local oscillator, ref and probe), all manufactured by OEWaves (OE4040-XLN) except for the probe, which was made by PurePhotonics (PPCL550). A broadband electro-optic frequency comb is derived from the local oscillator laser and serves as a frequency/phase reference to stabilize the remaining four lasers. Each laser is then modulated at the experimental clock rate of 1 MHz to provide temporal modes suitable to their purposes: the pump lasers are carved into 1-ns pulses (Exail MXER-LN-20, DR-VE-10-MO), the ref and probe lasers are carved into 400 ns (with AO Fiber pigtailed Pulse Picker from AA Opto-Electronic) and 50-ns pulses (using Exail MXER-LN-20), respectively, interleaved with the pump pulses at a later stage. The local oscillator laser has its complex temporal envelope mode-matched (using Exail MXIQER-LN-30) to the output of a representative squeezer (selection process is described in Supplementary Information). Each pulse train is amplified to a suitable power using erbium-doped fibre amplifiers (Model Pritel MC-PM-LNFA-20) before being combined into a polarization-maintaining fibre (with Opneti PMDWDM-1-1-CXX-900-5-0.3-FA). P1, P2 and local oscillator beams again have their phases stabilized, before being distributed among two sets of channels. The first channel, of which there are 24 copies, contains P1, P2, probe and ref, and is sent onwards to the sources array. The second channel, of which there are five copies, contains the local oscillator and ref, and is sent onwards to the QPU for coherent detection. Owing to the manner in which the beams are distributed, the ref laser carries the phase information in the system and thus can be used to stabilize measurement angles in the QPU. A second copy of the ref beam, slightly detuned in frequency, is sent on to the refinery array and used to stabilize interconnections between sources and refinery as well as between refinery and QPU.

Sources array

Each of the 24 sources chips are identical in design, and are based on the silicon-nitride waveguide platform provided by Ligentec SA and fabricated on their 200-mm production manufacturing line at X-Fab Silicon Foundries SE. Squeezers in these devices are based on a photonic molecule design, in which a pair of microring resonators are tuned to enable degenerate squeezed light to be generated using a dual-pump scheme, while leading-order unwanted parasitic nonlinear processes are suppressed[6]. The generated squeezed states are characterized using optical heterodyne tomography and found to be nearly single mode. The local oscillator temporal mode is matched to the dominant temporal mode as described in Supplementary Information. The outputs of the squeezers are passed through integrated asymmetric Mach–Zehnder interferometer (MZI) filters to remove pump light, then entangled by a tunable linear optical interferometer. Tuning of the interferometer, filters and resonators is accomplished using thermo-optic phase shifters. The chips are 8 mm × 5 mm in size, and are fully packaged and encased in a modular enclosure that mounts on a custom backplane chassis assembly. The end-to-end insertion loss (from pump input to quantum output) of the chips is 2.16 dB, and 1.82 dB of loss is experienced by the quantum light from when it is generated to when it is available in the fibre outputs. This figure includes an estimated squeezer resonator escape efficiency of (88 ± 3)%, filter and interferometer propagation loss of (0.36 ± 0.04) dB, and chip out-coupling loss of (0.90 ± 0.15) dB (81% coupling efficiency).

Newer designs based on different fabrication platforms and components, although not yet deployed in Aurora, have since been developed that combine multiple layers of thin silicon-nitride waveguides with a dispersion-optimized thicker layer, enabling much lower chip–fibre coupling loss, and higher squeezing through better suppression of parasitic nonlinear processes. Single-mode waveguide propagation losses of approximately 2.2 dB m^{-1} have been demonstrated, with even lower losses available in wider cross-sections for resonators. Escape efficiencies in squeezer microresonator structures exceeding 98% are routine, but further improvements in design and fabrication are needed to maintain an acceptable loaded quality factor under such strong over-coupling conditions. Similarly, chip–fibre coupling from these devices with losses of approximately 0.1 dB have been observed, with simulations for future structures indicating that arbitrarily low losses are possible. These results are consistent with recent progress reports in low-loss quantum photonic component development[26]. Even once realized, maintaining such low chip–fibre coupling losses through the packaging process in a manufacturing line capable of producing the millions of chips needed to furnish machines of practical utility remains an outstanding challenge.

PNR detection system

The PNR detection system is based on an array of 36 transition edge sensors (TES), housed in a pair of Bluefors (LD400) dilution refrigerators at 12-mK base temperature. These sensors are inductively coupled to an array of coherent superconducting quantum interference devices (SQUID) for cryogenic amplification, the signals from which are digitized and analysed in real time by an array of field-programmable gate array boards that discriminate photon number from the analogue pulses emerging from each sensor. Previously, such TES detectors were limited to repetition rates of a few hundred kilohertz, owing to their intrinsic thermal reset behaviour; higher experimental repetition rates required the use of demultiplexers[22], which are undesirable owing to their added loss and complexity. The TES detectors employed in Aurora enjoy native operation at 1-MHz repetition rate, while preserving photon miscategorization error below 10^{-2} for photon numbers up to 7. To enable this, the sensors were fabricated with small gold fins deposited at the margins of the tungsten absorber area; this engineers the thermal response of the detectors to absorbed photons to become faster by increasing the electron–phonon coupling with minimal impact on the other performance metrics of the detectors. The detection efficiency of this generation of TES detectors was not optimized for Aurora, and ranged from 97% to 69.3%.

The majority of the spread in this detection efficiency is believed to originate from variations in the detector packaging process. In Aurora, the deployed sensors were assembled by hand with no special quality assurance process enforced. More recently, TES detectors with operational speed at or above 1 MHz have been measured with consistently high detection efficiency above 90%. This is expected to increase to

at least 97%, matching the best channels available in Aurora, as a more reliable packaging and assembly process is developed. Still, PNR detection efficiencies over 99% are needed to stay within the loss budgets for the P_1 path in our architecture (Fig. 1). Our simulations indicate that the primary challenge to achieving this, apart from repeatable and reliable assembly processes, lies in obtaining tight process control over the multilayer dielectric stack parameters used to form the optical cavity that enhances photon absorption in the tungsten film.

Refinery array

The refinery array consists of 6 nominally identical PICs, 14.6 mm × 4.5 mm in size, based on the thin-film lithium-niobate PIC platform offered by HyperLight and fabricated on a semiconductor volume manufacturing line. Two binary trees—each composed of three electro-optic Mach–Zehnder modulator switches—select, based on feedforward instructions provided by the PNR detection system, the switch pattern that optimizes the output state. These MZI switches have an average insertion loss of 0.19 dB, giving an average total insertion loss of 4.15 dB for the full optical path through each refinery chip, which includes the binary tree, chip in- and out-coupling, and Bell pair entangling beamsplitter losses. Pairs of refinery chips, all of which are fully packaged, are hosted in three rack-mounted enclosures. Appropriate duty cycling of the switch settings between quantum pulse time windows allows the voltage bias of each modulator to be continually monitored and locked, yielding an average switch extinction ratio of more than 30 dB.

In Aurora, the refinery chips implement probability-boosting multiplexing as well as Bell pair synthesis, but do not have homodyne detectors at the multiplexer switch outputs and thus do not implement breeding. Efforts are underway to integrate photodiodes into the refinery chip platform, and we expect the next generation of refineries to implement the full adaptive breeding protocol.

Although the modulation and detection bandwidths demanded by the refinery's functions are not especially challenging compared with other applications, the loss requirements are. The importance of lower losses in MZI switches is compounded by the number of them present in the various optical paths. In the versions of the architecture contemplated in Fig. 4, the deepest combined path (P_2 and P_3 in Fig. 1) incorporates as many as 15 MZI switches. Even neglecting all other losses, this would mean that no more than approximately 7 mdB can be tolerated in each switch. Efforts to obtain this are well underway, with recent design and process optimizations yielding performance consistent with losses of 30 mdB per MZI switch. Early indications point to the importance of optimizing the thin-film lithium niobate (TFLN) etching process to manage scattering losses in the underlying waveguides. In addition, electrical control approaches must be engineered that enable high driving voltages. The MZI switch loss is nearly proportional to its length; shorter modulator sections are perfectly acceptable optically, but require proportionately higher applied voltages to operate. The driving approach must be scalable to allow operation of thousands of adaptive switches on the same chip. High-voltage-compatible integrated circuit fabrication nodes are being explored for this purpose. Other approaches using alternative materials such as barium titanate[26] have shown promise for delivering lower-voltage operation in this context, but further process improvements would be needed to compete with TFLN on raw propagation loss.

QPU array

The QPU array consists of 5 nominally identical modules, each based on a 300-mm silicon photonic-chip platform offered by AIM Photonics that hosts silicon-nitride and silicon waveguides, germanium photodiodes, and carrier depletion modulators. Within the chips, each measuring 6.2 mm × 4.3 mm and fully packaged within a rack-mounted enclosure, silicon nitride is used for edge coupling from the fibre inputs and for the interferometer that implements spatial entanglement in the cluster state. The quantum light then transitions to silicon waveguides and is mixed with local oscillator light on appropriate beamsplitters, and terminates on germanium photodiodes for homodyne detection. The loss experienced by each quantum input to the QPU is on average 3.68 dB, of which 0.82 dB arises from the edge couplers and optical packaging, 2 dB from the interferometer circuit and 0.86 dB from the photodiodes. The local oscillator input to one homodyne detector is modulated using silicon carrier injection modulators, with each quadrature phase setting actuated based on real-time instructions from the digital QPU controller. This controller is based on a field-programmable gate array, which is programmed to select the appropriate measurement bases, taking into account the algorithm and decoder protocol selected by the user.

The full signal chain latency from an optical pulse arriving at the homodyne detectors to the actuation of an updated local oscillator phase is approximately 976 ns. Out of this, 240 ns is spent on converting the optical homodyne pulse to a normalized 16-bit fixed-point number, involving photodiode response, transimpedance amplification, analogue-to-digital conversion, and digital signal processing for pulse integration and normalization. Another 672 ns is the worst-case serial link latency spent on serialization, propagation to the QPU backend, de-serialization of the 16-bit number, plus serialization, propagation back to each of the 5 QPUs, and de-serialization of the 2-bit local oscillator phase selection command. The remaining 64 ns are used for the decoder algorithm to calculate the next local oscillator phase setting from measurement information of all homodyne detectors from previous clock cycles. Future decoders requiring more digital clock cycles to carry out intervening computations could be deployed on digital circuits with higher clock speeds, or the increased latency could be offset by latency improvements in other parts of the signal chain. For example, latency could be improved by selecting lower-latency analogue-to-digital converters for digitizing the homodyne measurements, by optimizing the signal processing chain for homodyne value normalization, and by improving the serialization and de-serialization latencies associated with the serial links by means of using field-programmable gate arrays with high-speed (>1 Gbps) serial input/output pins for the QPUs.

The full fault-tolerant design, involving refineries that incorporate breeding, would have identical chip platform requirements for the refinery as for the QPU, those being electro-optic modulators and photodiodes. In the future, we thus expect both the refinery and the QPU to be based on the same PIC platform, which will look closer to the TFLN-based substrate used for the refinery in this work.

Our recent work optimizing the design of integrated photodiodes has yielded quantum efficiencies as high as 98.5% in the same germanium platform as that used in Aurora; the heterogeneous integration of high-quantum-efficiency photodiodes like these into TFLN devices is the last platform integration step required to equip Aurora with refineries capable of implementing breeding protocols. Considering this co-integration requirement with TFLN, our focus has turned towards using III–V-semiconductor-based photodiodes in place of germanium. Simulations indicate that evanescent coupling between TFLN waveguides and InGaAs photodiodes, appropriately fabricated, can deliver homodyne detectors with net quantum efficiency well above 99%. The most challenging aspect of achieving this lies in the details of the heterogeneous integration scheme used. High-quality surface preparations within deep trenches will be needed at the interface between the photodiodes and the waveguide cladding. Managing excessive dark current in the photodiodes themselves as their dimensions grow to accommodate near-unity absorption will also require innovative approaches.

Interconnects

Between the sources and refinery modules, and between the refinery and QPU modules, interconnects are required that can provide a specified and fixed delay on the quantum pulses conveyed. For the sources-to-refinery links, the delay serves as a buffer for awaiting heralding information from the PNR detectors, whereas the refinery-to-QPU links implement

a delay of exactly one clock period on half of two different Bell pairs, enabling temporal entanglement in the cluster. The purposes of these delays differ slightly but they are otherwise identical in requirements. In particular, both must actively stabilize the link against fluctuating phase and polarization. This is accomplished by interleaving coherent classical reference pulses between each quantum pulse, interfering reference pulses between appropriate inputs to each chip, and feeding back on phase and polarization actuators in the fibre delay modules.

The delay lines themselves are implemented in discrete enclosed modules within the racks, and are each composed of a fibre coil of (253.286 ± 0.009) m (about 1.239 μs). It is noted that the fibre coils in 10 out of the 12 channels between refinery and QPU are (48.762 ± 0.004) m long (about 0.239 μs) such that the difference between the two interfering channel is exactly 1 μs. Each fibre coil is in thermal contact with a thermo-electric cooler, which provides slow phase tuning with a large capture range, compensating for phase drifts that arise from the unstable global temperature environment in the racks. These are accompanied by piezoelectric fibre phase shifters (Luna FPS-001) that provide fast phase control over a smaller range, locking against acoustic fluctuations. An electrical polarization controller (Luna MPC-3X) within each fibre module is used to ensure the inputs to each chip are aligned to the appropriate waveguide mode.

These custom first-generation modules were manufactured by Luna Innovations, and have an average loss of (0.28 ± 0.08) dB, excluding connectors, while adding between 0.6° and 1.5° of phase noise (root mean square), when in closed loop operation.

Recent prototype designs have demonstrated losses of <0.1 dB, of which 0.037 dB arises simply from the length of fibre itself in the delay. Future generations of this module are expected to be limited by only the fundamental propagation loss in the underlying fibre, which can be as low as 0.14 dB km^{-1} in existing, commercially available products. Our initial studies have shown that the vast majority of the loss present in typical fibre delays arise from fibre connectors or splices between different constituent components. These are straightforward to eliminate by manufacturing fibre delay modules from a single draw of fibre. Thus, even without increasing the quantum clock speed beyond 1 MHz, using ultralow loss fibre and implementing these manufacturing changes is expected to achieve about 0.03 dB (0.7% loss). Going beyond this would require faster clock speeds or lower loss fibre. In addition, the inherent modularity of the architecture allows for all the fibre interconnections to be spliced, or indeed discrete fibre components in a given off-chip path to be assembled from a single draw of fibre.

Entanglement verification

To verify the generation of multimode entanglement, we investigate the nullifier variance[48] of the six Bell pair input states when the refineries are set to deterministically output squeezed states (that is, each pair is an approximate two-mode squeezed state). As we are interested in both q and p correlations, we alternate the cluster-state acquisition between the two measurement bases, measuring all modes in q and p, subsequently. Experimentally, the measurement basis was changed by a simultaneous phase rotation of all input modes using a setpoint change at our arbitrary-phase locks. For each basis, the data are acquired continuously over 2 hours, amounting to an uninterrupted measurement of 7.2 billion time bins (yielding 86.4 billion modes in total, combining all 12 operating modes at each time step). For the statistical analysis described below, the acquired data are processed in batches of 10 million time bins.

In the analysis, we obtain the statistical moments of the individual Bell pair states by 'reverting' the cluster-state stitch at the QPU (also known as synthesis of the macronode[7]). As a first step, we build the quadrature covariance matrix Γ of all measurement outcomes. The elements of the covariance matrix are given by $\Gamma_{ij} = E(x_i x_j) - E(x_i)E(x_j)$, where a subscript represents a spatiotemporal mode and the operator E denotes the expected value (mean) of its argument. Our covariance matrix is of dimension 24 × 24 as we aim to capture not only the correlations among the 12 spatial modes but also their correlations to the quadratures of the subsequent time bin (12 modes for time bin t plus 12 modes for time bin $t+1$). For this demonstration, we either measure all modes in $\hat{q}$ or all modes in $\hat{p}$, so we can build only the position–position or momentum–momentum subblocks of the full covariance matrix, which is sufficient for evaluation of Einstein–Podolsky–Rosen (EPR)-state nullifiers.

As part of the macronode synthesis, the six input Bell pair states are stitched together by a beamsplitter network, represented by the symplectic transformation S. In our second analysis step, we apply the inverse of that symplectic to our quadrature covariance matrix to obtain the covariance matrix before the macronode stitch: $\Gamma_{\text{in}} = S^{\text{T}}\Gamma S$. This back-transformed covariance matrix now represents six separable EPR states. Their nullifiers are defined as $\hat{n}_q = (\hat{q}_0 - \hat{q}_1)/\sqrt{2}$ and $\hat{n}_p = (\hat{p}_0 + \hat{p}_1)/\sqrt{2}$. The variance of these nullifiers is obtained using the definition $V(x_i \pm x_j) = V(x_i) + V(x_j) \pm 2\text{cov}(x_i, x_j)$, where all variance and covariance terms are obtained from the elements of Γ_{in}. To obtain the shot-noise reference for our EPR nullifiers, we apply the same operation to the vacuum data, obtained with all squeezed-light sources turned off.

It is noted that our method to evaluate the EPR nullifiers by applying the inverse macronode symplectic S^{T} to the output covariance matrix Γ is mathematically equivalent to applying S to the nullifier equations $\hat{n}_q$ and $\hat{n}_p$ and evaluating the resulting equations using Γ directly.

Adaptivity demonstration

In the adaptivity demonstration, we created the set-up for a distance-2 repetition code implemented through a cluster state composed of low-quality GKP states and squeezed states. Although the system is too noisy to show error suppression, we nevertheless demonstrate all the building blocks, collecting and processing of measurement data, running a decoder in real time, and performing a conditional operation at the following time step based on the recovery. The demonstration can be broken down into the following steps (Extended Data Fig. 2), with additional detail supplied in Supplementary Information section VII D:

1. Initialization. In time step τ, homodyne measurements decouple qubits M_0, M_1, and M_2 ($\hat{q}$ measurements on all modes) and teleport qubits M_3, and M_4 ($\hat{p}$ measurements on modes 3 and 10, and $\hat{q}$ elsewhere) to initialize the experiment.
2. Memory measurement. In time steps 2τ to 4τ, homodyne measurements are performed corresponding to two foliated repetition code checks (qubits M_0, M_3, and M_4 measured in X, that is, $\hat{p}$ measurements on modes 3 and 7, and $\hat{q}$ measurements elsewhere) and accompanying teleportations (qubits M_1 and M_2 measured in X, that is, $\hat{p}$ measurements on modes 4, 9 and $\hat{q}$ measurements elsewhere).
3. Decoding. All decoding occurs in time step 4τ, as follows:
   (i) Inner decoding. The raw homodyne measurement outcomes from the previous steps, along with the state record (from photon-counting outcomes), are processed to obtain bit values and the probability of qubit error.
   (ii) Outer decoding. The syndrome and qubit error probability are passed to a qubit-level decoder. After a few iterations of the decoding algorithm (belief propagation), the decoder outputs the recovery along with updated estimates of error probabilities.
   (iii) Decision. A thresholding function is computed on the decoding output to assess confidence in the recovery. For a pre-determined threshold, we decide whether to 'keep' the entanglement and perform an additional repetition code check or to 'cut' the entanglement and restart the experiment. Steps i–iii constitute a complete real-time decoding round. In the control experiment (without feedforward), the decision bit (0 for 'cut' and 1 for 'keep') is added to a random bit, decoupling the decoding from the feedforward action.
4. Feedforward and adaptive measurement. The decision based on the recovery obtained in time step 4τ is transmitted to qubit M_0 in the next clock cycle, at time step 5τ. The local oscillator phase of

mode 7 in this qubit is changed accordingly. In the 'keep' case, this qubit is measured in X (mode 7 in $\hat{p}$), preserving the entanglement, and in the 'cut' case it's measured in Z (mode 7 in $\hat{q}$), cutting the entanglement. Mode 6 is always measured in $\hat{q}$. Qubits M_1 and M_2 are measured in Z (all modes in $\hat{q}$).

5. Reset. Finally, also at time step 5τ, qubits M_3 and M_4 are decoupled through Z measurements (all modes measured in $\hat{q}$) to reinitialize the experiment.

To confirm that the correct measurement has been performed, we plot correlations between four modes in the same clock cycle. In the 'cut' case, ideally no correlations should be observed, whereas in the 'keep' case, we expect to see cross-correlations between modes.

Data for the decoding demonstration are acquired in 69 batches of 1 million time bins each, amounting to a total of 69 million time bins and 13.8 million repetitions of the decoding algorithm. In the post analysis, the quadratures acquired in the final decoder time bin are separated into two groups, one for each of the two decoder-determined outcomes X and Z. Instances in which one or more cat inputs are involved in pairs 3 and 4 are discarded from the analysis. For both groups, we build a 12×12 covariance matrix of the quadratures acquired in the adaptive time step (corresponding to the central M_0 macronode and its neighbours). Classical cross-correlations among quadratures are obtained in a separate vacuum measurement (with squeezed-light sources disabled) and their covariance matrix is subtracted from the X and Z covariance matrices. Experimental covariance results are compared with theoretical predictions obtained from circuit simulations assuming that all modes were squeezed states with 4 dB of initial squeezing and 5% total efficiency (including optical loss and mode matching).

Optimizing candidate state factories for loss tolerance

We optimize over configurations of elements that produce GKP Bell pairs to be entangled into a cluster state. For the choice of cluster state, we select the Raussendorf–Harrington–Goyal cluster-state lattice[49–51]. We use a two-layer decoder that first obtains a syndrome compatible with the ideal GKP qubit subspace, followed by a minimum-weight perfect matching algorithm[52] to find a recovery operation that minimizes the probability of a logical error. The first 'inner' layer of this decoder scheme takes into account the correlations present in the noise arising both from the probabilistic nature of the state-generation process and from the extra modes present within each macronode and its neighbours (see ref. 37 for more details). The second 'outer decoder' is also informed by a set of marginalized probabilities of errors furnished by the inner one. This strategy yields an effective squeezing threshold for fault tolerance of 9.75 dB, which improves on the 10.1-dB value found in ref. 29. The quantum error correction (QEC) squeezing threshold defines the quality of the GKP states that must be available, which then determines the loss requirements for each path. The bound of the loss requirements for these paths can be found by searching different architecture configurations with loss included along each optical path. Full details of the calculations behind the effective squeezing metric and corresponding fault-tolerance thresholds are available in Supplementary Information.

The GKP-state factories considered in this work are parameterized by the GBS device settings (number of modes, levels of squeezing, interferometer angles, maximum detectable number of photons, loss levels), the refinery settings (depth of the binary tree, ranking function/selection rule of the refinery inputs, the number of refinery inputs to undergo breeding, degree of measurement-based squeezing, loss), and the QPU switch tree depth and associated loss. To evaluate a candidate state factory, we perform a Monte Carlo simulation to sample the distribution of output states over the PNR statistics of the GBS devices and the homodyne statistics of the refinery. From the distribution of states, we can determine the quality (symmetric effective squeezing) of the average Bell pair, taking into account the loss it will experience. (It is noted that that the symmetric effective squeezing of the refinery outputs is distinct from the amount of squeezing assumed in the GBS cells, which we fix to be 15 dB.) We leave P_1, P_2, and P_3 as free parameters but fix the P_4 path to be 40–50% of P_2 depending on the refinery depth as those paths have similar elements (the squeezers, chip input/output, fibres, homodyne detectors) and differ only by the binary tree of adaptive beamsplitters used for MUX and breeding. For more details, see Supplementary Information. The quality of the average state can then be compared with the 9.75-dB result of threshold calculations for the choice of cluster state and decoder to determine whether such states would be suitable for fault tolerance. Despite the states being highly non-Gaussian mixed states, we can simulate thousands of output-state samples from a single factory in a few minutes on a single core, and explore tens of thousands of factory candidates in reasonable time using a computing cluster. To achieve this, we employ different quantum optical representations at different stages of the state factory, namely, the Fock, Bargmann, characteristic and quadrature basis pictures, depending on what needs to be computed at that point in the factory (PNR statistics, GKP stabilizer expectation values, beamsplitter interactions, homodyne statistics). This is numerically implemented using MrMustard[53], an open source software package for simulating and optimizing quantum optics circuits.

Data availability

The datasets generated and analysed for this study are available at https://github.com/XanaduAI/xanadu-aurora-data.

48. Menicucci, N. C., Flammia, S. T. & van Loock, P. Graphical calculus for Gaussian pure states. *Phys. Rev. A* **83**, 042335 (2011).
49. Raussendorf, R., Bravyi, S. & Harrington, J. Long-range quantum entanglement in noisy cluster states. *Phys. Rev. A* **71**, 062313 (2005).
50. Raussendorf, R., Harrington, J. & Goyal, K. A fault-tolerant one-way quantum computer. *Ann. Phys.* **321**, 2242–2270 (2006).
51. Raussendorf, R., Harrington, J. & Goyal, K. Topological fault-tolerance in cluster state quantum computation. *New J. Phys.* **9**, 199 (2007).
52. Dennis, E., Kitaev, A., Landahl, A. & Preskill, J. Topological quantum memory. *J. Math. Phys.* **43**, 4452–4505 (2002).
53. XanaduAI. MrMustard. *GitHub* https://github.com/XanaduAI/MrMustard (2021).

Acknowledgements We acknowledge A. Lukashchuk and Z. Zaidi for assistance with module assembly; A. Daniel, D. Herr, T. Matsuura, G. Pantaleoni, E. Sabo and H. Yamasaki for theory assistance; A. Goussev, G. Karelov, R. Lasaten, R. Saber and Y. Yurchenko for laboratory operational support; and J. M. Arrazola and C. Weedbrook for feedback on the paper.

Author contributions R.B., H.C., P.C., I.D.L., P.E., T.G., I.K., E.L., C.M., M.M., K.R.S., S.A.S. and X.X. designed, developed and characterized the packaged PICs used in Aurora, supervised by B.M. S.I., S.K., J.F.T., J.E.T., V.D.V. and Y.Z. designed and built the modular pump, sources, refinery and QPU systems, supervised by D.H.M. H.A.R., N.D'A., S.E.F., D.S.P., F.R., J.S.-C., Q.Y.T. and R.B.T. developed and built the PNR detection system, supervised by M.J.C. T.A., B.A., D.D., L.G.H., J. Hundal, M. Seymour and M.Z.A. developed the software control system, supervised by L.N. I.C., M.J., P.L., M.L., B.L., K.P., S.S.-B. and E.S. designed the electronics and mechanics for the modules, supervised by M. Stephan M.F.A., L.S.M. and X.Y. developed the phase stabilization system and systems integration strategy, and participated in data taking for the experiment benchmarks alongside F.L., who also carried out the systems initialization, simulation and data analysis. T.N. and L.F.S.S.M.V. developed the fibre delay lines, supervised by J.L., who also supervised the systems integration and experimental benchmarking. J.E.B., G.D., H.F., J.G., M.V.L., F.L., S.D., F.M.M., P.J.N., T.R., M. Seymour, M. Silverman, I.T., M.V. and B.W.W. developed the theoretical details of Aurora. L.B., R.S.C., R.D.P., S.F., S.G., C.G.-A., J. Hastrup, T.H., C.E.L., A.P.L., S.D., F.M.M., Z.N., R.N., W.N.P., T.R., M. Silverman, Y.S.T. and Y.Y. developed the GKP-state preparation theory used, supervised by J.E.B. B.Q.B., L.M.C., H.F., J.G., C.G.-A., Z.H., T.H., T.J., A.P., T.R., M. Silverman, M.V., B.W.W. and R.W. developed the measurement-based quantum computing theory used, supervised by I.T. T.H., P.J.N., A.P. and M.V. developed the quantum error correction theory used, supervised by G.D. Development of all theory used was supervised by R.N.A. Z.V. supervised the project, and co-wrote the paper with R.N.A. and J.L., with input from all co-authors.

Competing interests The authors declare no competing interests.

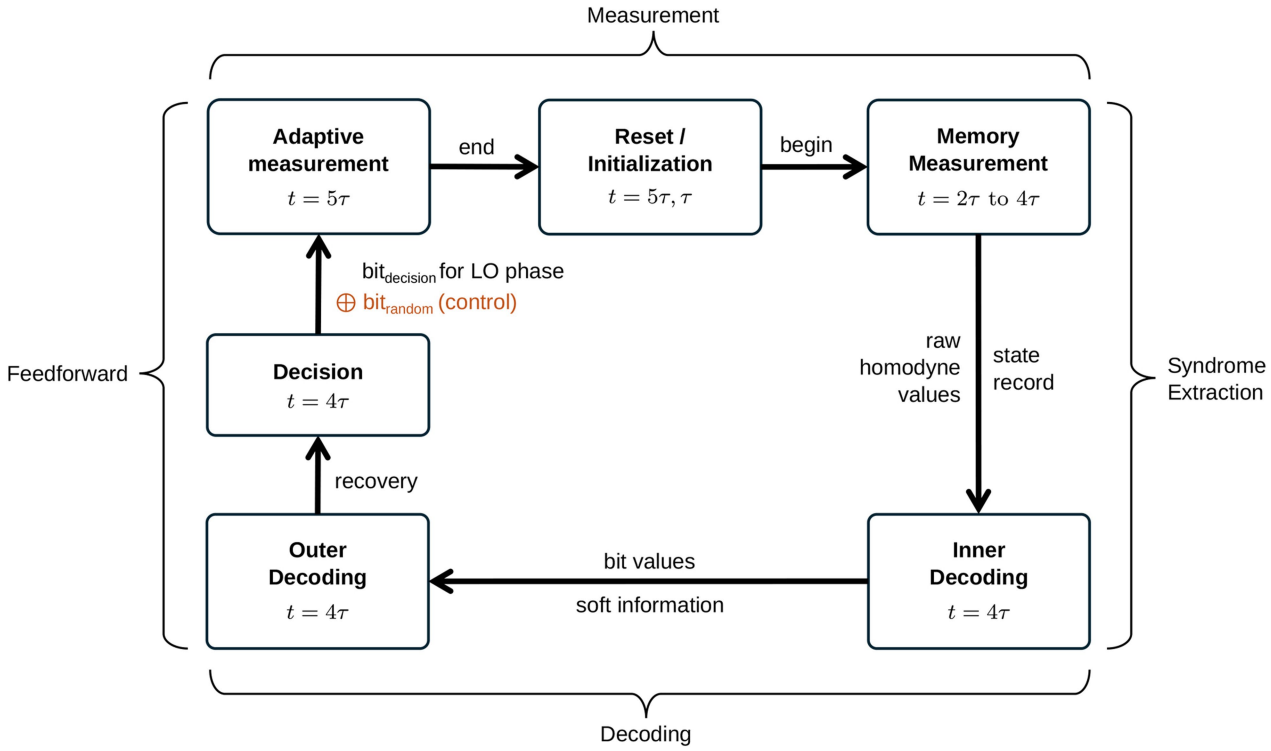
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08406-9.

Correspondence and requests for materials should be addressed to R. N. Alexander or J. Lavoie.

Peer review information *Nature* thanks Niklas Budinger, Jonas Neergaard-Nielsen, Olga Solodovnikova and the other, anonymous, reviewer(s) for their contribution to the peer review of this work

Extended Data Fig. 1 | Photograph of the Aurora system. The entire system, apart from the cryogenic detection array, fits into four standard 19-inch server racks and is fully operated using a single server computer.



Extended Data Fig. 2 | Summary of the steps in the adaptivity demonstration. The cycle begins at the middle block of the measurement step, when the cluster state is initialized. Repetition code stabilizer measurements and teleportations are performed in time steps 2 to 4. The decoder acts on data collected in these time steps and makes a decision about a measurement to perform in the following time step. In the control experiment, this decision is randomized. On the fifth and final time step, the measurement basis is modified based on this decision.