Network Security

SL5 Novel Recommendations

NOVEMBER 2025

Peter Wagstaff

Lisa Thiergart

Yoav Tzfati

Luis Cosio

Guy

Philip Reiner

Overview

Executive Summary

In the field of network security, we are fortunate to have many established solutions to build on, from both industry and the national security ecosystem. However, these solutions come with major trade-offs and will need to be adapted to the unique and exceptional performance requirements of AI workflows. Our recommendations include **strong network perimeter security** to keep attackers out (and AIs in). They also call for **stronger encryption standards in key areas**. A major source of tension is the practical need for resources to be geographically distributed or remotely accessed and the difficulty of securing long-distance connections. Whether long-distance communications can be secured remains uncertain. **Trade-offs exist between trusting long-distance communications and the extent to which operations can be distributed.** Network design needs to **consider the transformative role automated AI R&D will play** in the future, and provide AI agents access to resources needed for running experiments through controlled and limited interfaces. In this report we share our top five of many recommendations for securing the networks enabling frontier AI development from nation-state level attacks.

Top 5 Recommendations

1. Connect network enclaves with inline network encryptors
2. Encrypt accelerator interconnect
3. Isolate powerful models on an air-gapped network
4. Maintain an additional air-gapped network for running unsigned code for R&D
5. Develop AI-enhanced cross-domain solutions

Connect network enclaves with inline network encryptors

Explanation of Recommendation

We recommend labs encrypt all traffic exiting a network enclave to travel over a less secure connection using an inline network encryptor. A network enclave could be a data center building, a secure area in an office building, or any other space where a high degree of physical security can be established. By contrast, long-distance connections between these may be infeasible to physically secure. Depending on how much confidence is placed in network encryptors, it may be necessary to avoid using these connections for the most sensitive communications, especially those involving weights such as distributed training.

Research Reasoning

By current trends, training a frontier model may require over 5 GW of power by 2029 [1]. For reference, this is almost half the peak demand of all of NYC [2], and more than any power plant in the US except the Grand Coulee Dam [3]. The compute required for a single frontier AI lab may exceed 10 GW [4]. It is unlikely that this much demand can be supplied at a single location, barring an unprecedented effort in power plant construction. Therefore, it is likely SL5 infrastructure will be geographically distributed, requiring connections to be made over significant distances. These connections must be secured.

USG classified networks like SIPRNet and JWICS overcome this problem using inline network encryptors [5]. These high assurance devices encrypt traffic as it leaves a classified enclave and travels over a lower security connection. USG networks and their contractors use NSA Certified Type 1 encryptors, which are not available for normal commercial use [6] and require strict handling requirements [7]. Alternatively, enterprise network encryptors can offer a similar level of security with many of the same features, complying with FIPS 140-3 Level 3 [8]. In our discussions with experts, we were told that properly used, they may be just as good.

We heard claims from at least one expert that while the cryptographic primitives that underlie network encryptors are highly unlikely to have vulnerabilities, there might exist vulnerabilities in the implementation that could be exploited by a highly capable attacker. In this case, it may be possible to **guard against implementation flaws specific to a single model by using multiple, heterogeneous network encryptors from different suppliers in series**. This is consistent with the NSA's "Rule of Two" [9]. Determining the true vulnerability of these systems may require access to classified information; we leave open the possibility there may be no way to securely make long-distance connections. If these connections cannot be fully trusted, labs may wish to make the most **sensitive transfers with physical storage medium**, if at all.

Using network encryptors may present some implementation challenges. Both NSA certified and enterprise encryptors currently offer up to 100-400 Gbps of bandwidth [10], [11]. If training needs to be distributed, future distributed training runs may require over a hundred times more bandwidth [12]. Due to slow development and certification cycles, a solution **may need to rely on multiplexing currently available encryptors**. This is possible, but to date has only been done with small numbers of encryptors [10], likely because there has not yet been a need for more. In **current inter-data center connections, each fiber only carries ~400 Gbps regardless of the total bandwidth** [13]. There may be little need for additional multiplexing if each fiber can be handled by one encryptor. It is also possible that other advancements allow efficient distributed training with much less bandwidth [14], or that bandwidth-intensive tasks can be collocated. This would be desirable since bandwidth-intensive tasks usually involve weights directly.

Cost-Benefit

- **Costs:** Standard pricing is not available, but high bandwidth network encryptors could cost well over $100k each [15]. In scenarios where this supports high bandwidth connections between locations, the size of this demand may also strain the current supply. Multiplexing may present additional costs and challenges. **Latency impact is negligible—on the order of ~2 µs, trivial compared to end-to-end latency** [8], [11].
- **Benefits:** Provides the level of security nation-states depend upon for connections made over untrusted network segments, enabling connections to multiple locations, including office buildings. Adds an additional layer of defense against an insider threat sending unencrypted (or decryptable) traffic outside a network enclave.
- **Alternative:** Rely on end-to-end encryption for all communications on the network (note that you should have this anyway). This alternative is likely unacceptably vulnerable, especially to insider threats. Alternatively, make the entire network exist in a single building. This creates a power bottleneck and forces your employees to relocate there.

Encrypt accelerator interconnect

Explanation of Recommendation

We recommend labs encrypt communications over AI accelerator inter-chip interconnect. Messages should be **encrypted on the sending accelerator and decrypted on the receiving accelerator**. This likely requires chip designers to include new features in accelerators, which need to be planned in advance to account for long lead times. Some efforts may already be underway to prepare for this, and we encourage **expanding this effort to all relevant AI chips including novel types that may be in use in the years 2027 and onward**.

Research Reasoning

Broadly, weights and values from which weights could be easily derived should never be stored or transmitted in cleartext. While this is the ideal, one area where this is challenging to implement is accelerator interconnect, which transmits weights and weights related values between accelerators with tremendous speed and volume [16]. The additional latency and compute overhead of encrypting these messages before they leave an accelerator could be impractical if not implemented efficiently [17].

Encrypting interconnect may require dedicated hardware on accelerators, which is not yet standard. Accelerators are optimized to perform only certain operations, which do not include operations crucial to encryption (especially stream ciphers) like XOR [18]–[20]. NVIDIA Hopper GPUs do not have dedicated hardware encryption for GPU-GPU communications over NVLink [21]; however, their new Blackwell GPUs do [22]. Trainium does not currently have dedicated encryption hardware for its NeuronLink interconnect [23], nor does TPU [24]. Future accelerators have been rumored to include these features according to experts we talked to and as previously stated, we strongly encourage the expansion of this to all relevant AI chips.

Since there is a significant delay between chip design and availability, it is **necessary for these features to be requested and planned well in advance**.

Given the difficulty of encrypting interconnect, alternative mitigations could be explored. An attacker intercepting interconnect traffic would require physical access to the cables, which can be physically protected. Currently, these cables benefit from the security of being inside a data

center but are not themselves protected. They could be **secured with an additional enclosure, potentially including tamper-proof or tamper-evident features**.

Cost-Benefit

- **Costs:** Including hardware for encryption will increase the cost of accelerator chips, but once the chip makers decide to include it, that cost becomes built-in. For overhead, NVIDIA claims their solution offers "nearly identical throughput" [22].
- **Benefits:** Interconnect no longer carries sensitive cleartext traffic.
- **Alternative:** Continue to send sensitive cleartext traffic over interconnect and instead rely on enhanced physical security, such as tamper-proof or tamper-evident enclosures. This could make maintenance much more difficult since it needs to be somehow distinguished from tampering.

Isolate powerful models on an air-gapped network

Explanation of Recommendation

We recommend that frontier labs train and host frontier models that pose a catastrophic risk exclusively on an air-gapped network. Only signed and trusted code should be executable on this network. Serve customer inference with weaker models (or models post-release that clear necessary safety & security thresholds) at SL4.

Research Reasoning

Connections to the internet present a large attack surface, so systems requiring the highest security are often isolated from it [25], [26]. This separation is best enforced by a physical air-gap. Software separation, while convenient, is less reliable. Systems relying on air gapping include USG classified networks like SIPRNet and JWICS as well as critical infrastructure like nuclear power plants or financial systems [27].

A major vulnerability of an isolated network is the software brought onto it. It must be thoroughly tested and reviewed before being imported [27]. For example, requirements for SIPRNet prevent the installation of "any software and firmware components" lacking a trusted signature [28], [29], and also prevent the execution of unsigned mobile code [28], [30].

Air gapping should never create a false sense of security, and Zero Trust architecture is still desired since breaches are almost inevitable. This is consistent with the strategy networks like SIPRNet have pursued [31].

One may wonder what the point of creating a model is if it is fully isolated from the outside world. Even if no direct connections can be made, a highly capable AI could still provide tremendous value: Groundbreaking research or inventions could be published via the force of secure automated AI R&D. Weaker, more trusted models could be distilled and offered publicly. Future progress in AI Safety and AI security may even make air-gapping unnecessary in the long term. **SL5 is about securing the critical span when models become increasingly capable of automated AI R&D whilst AI alignment has not yet been sufficiently solved. During this critical window, their theft or sabotage poses untenable safety, security and national security risks which must be**

mitigated with SL5 and other governance and safety mechanisms. SL5 is a powerful tool to buy time to find and build more long-time tenable solutions to the AI alignment and governance problem than currently exist.

Cost-Benefit

- **Costs:** Air-gapping prevents automatic software updates, powerful models are restricted to internal use only, compute used to serve public inference cannot be used for training runs. Isolated models can still indirectly serve the public.
- **Benefits:** Large reduction in attack surface
- **Alternative:** Forgo air-gapping and focus on software isolation and/or zero trust architecture. Some industry leaders like Google have abandoned relying on perimeter security almost entirely in favor of zero trust architecture [32].

Maintain an additional air-gapped network for running unsigned code for R&D

Explanation of Recommendation

We recommend labs maintain at least two isolated networks:

- **Internal Network**: Trains and runs powerful models only available internally. Air-gapped and only runs trusted, signed code. Accessible from secure work areas in offices.
- **Development Network**: Enables running unsigned code for R&D purposes. Air-gapped and accessible from secure work areas in offices.

Research Reasoning (see Appendix A as well)

Software R&D workflows require running unsigned code. At the very least, writing and testing code requires running it, and it cannot be signed before it has been written and tested. Therefore, software development, test, and review require an additional network with the capability to run unsigned code. However, the software developed to create AI models is also highly sensitive. While it is not as directly useful as the models themselves, it contains critical knowledge about how to train models, which an adversary with their own compute could recreate. The Development Network also needs privileged access to the Internal Network to provide signed software to it and enable automated AI R&D. A compromise in the Development Network could lead to a compromise in the Internal Network. Therefore, the Development Network must also be air-gapped to achieve a similarly high level of security.

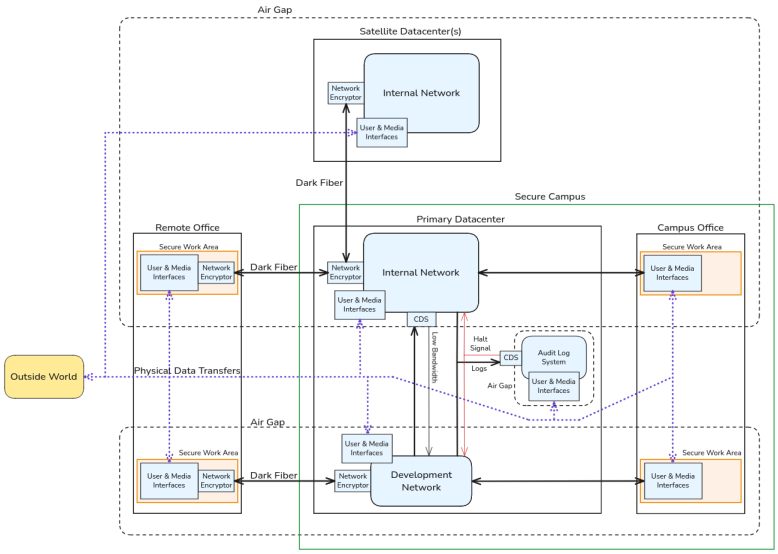


Fig. 1. High-level concept for SL5 Network, see Appendix A for more

Cost-Benefit

- **Costs:** A second network adds cost and complexity and requires a system to handle communications between them.
- **Benefits:** Creates separation between untrusted code and most sensitive operations. Enables a reduction in the software stack needed in weight storing systems (See Supply Chain recommendations).
- **Alternative:** Have a single network and rely on sandboxing to run untrusted code.

Develop AI-enhanced cross-domain solutions

Explanation of Recommendation

We recommend labs develop cross-domain solutions (CDS) enhanced with AI monitors to enable limited communication between the Internal Network and Development Network.

Research Reasoning

AI assisted or automated research workflows will likely involve thousands of models conducting experiments around the clock involving writing and executing code [33], [34]. If these AI models exist on the Internal Network where only vetted and signed code can be run, there must be a cross-domain solution to transfer this code and associated commands to the Development Network where unsigned code can be executed. Execution of this unsigned code should also occur in a sandbox environment, but these are not free of vulnerabilities [35], including powerful sandbox techniques used by cloud providers [36]. A capable AI may be able to exploit these. Additionally, this channel could be used to exfiltrate or distill the model weights into an environment where unsigned code could operate on them. A cross-domain solution is needed that addresses these concerns.

Cross-domain solutions are used, including in classified national security environments, to enable communications between two otherwise air-gapped security domains [37]. Existing cross-domain transfer solutions rely heavily on labels [37]. Information is associated with a cryptographically bound label that indicates among other things its classification level. The CDS can then apply rule-based functions to these labels to determine if the information can pass [38]. While this approach may be suited to sharing classified documents, it does not translate well to the task of newly written software and commands. In current implementations, labels are created through a tedious manual process [39]. This is not feasible for content newly generated by AIs at scale.

Additionally, the goal of labeling differs. “Classified information” can be analogized to weights and weight-related values which should never be exfiltrated to the Development Network, but the CDS should also reject malicious code and commands. While there may still be a place for labels and rule-based functions, to cover the full scope of the problem likely requires using an AI monitor to detect and reject suspicious files and commands. Relevant ideas from the space of AI control may be applicable [40]–[42]. Ideas in this space rely on weaker, more trusted AI models checking the outputs of more powerful AIs for suspicious activity, and involving human monitors in an efficient and targeted manner. More work should be done to research and develop these techniques.

Other cross-domain solution properties can be achieved with existing solutions. One-way connections can be enforced with data diodes [43], including variations that allow for packet acknowledgements and rate limiting [44].

Combining these features can ensure that cross-domain connections do not enable a high-bandwidth path from the Internal Network to the Development Network, defending against exfiltration of model weights. Bandwidth limiting is already a method being implemented to prevent model exfiltration [45]. Its efficacy is enabled by the large size of models (GPT-4 is believed to be ~3.4 TB [46]) in comparison to the volume of legitimate traffic (1 billion lines of code would be ~50 GB at 50 characters per line, for context all Google services are 2 billion lines [47]). Therefore, total egress can be kept below some safe fraction of model size.

Cost-Benefit

- **Costs:** AI monitors could consume a non-trivial amount of compute. These monitors should primarily be needed for connections leaving the Internal Network, since that is where suspicious AI actions would originate. This would mean these resources are still available for large training runs.
- **Benefits:** Enables AI assisted research and development workflows, while defending against attempts to execute malicious unsigned code or exfiltrate weights.
- **Alternative**: Restrict AI code and action space to what can be defined as safe by a set of explicit rules. This may be very restrictive. Place greater trust in sandboxing implementation.

Bibliography

[1] K. F. Pilz, Y. Mahmood, and L. Heim, “AI’s Power Requirements Under Exponential Growth: Extrapolating AI Data Center Power Demand and Assessing Its Potential Impact on U.S. Competitiveness,” Rand.org, Jan. 28, 2025. https://www.rand.org/pubs/research_reports/RRA3572-1.html

[2] New York Independent System Operator (NYISO), 2023 Load & Capacity Data Report (Gold Book), Table I-3a, p. 27, Apr. 2023. [Online]. Available: https://www.nyiso.com/documents/20142/2226333/2023-Gold-Book-Public.pdf

[3] “Electricity generation, capacity, and sales in the United States - top 10 - U.S. Energy Information Administration (EIA),” www.eia.gov, Oct. 26, 2023. https://www.eia.gov/energyexplained/electricity/electricity-in-the-us-top-10.php

[4] R. Dean, “Compute Forecast — AI 2027,” Ai-2027.com, Apr. 2025. https://ai-2027.com/research/compute-forecast#power-requirements

[5] J. Loyall, K. Rohloff, P. Pal, and M. Atighetchi, “A Survey of Security Concepts for Common Operating Environments,” IEEE, Mar. 2011, doi: https://doi.org/10.1109/isorcw.2011.31.

[6] “CSfC vs Type 1 Encryption: An Overview - Crystal Group,” Crystal Group, May 31, 2024. https://www.crystalrugged.com/knowledge/csfc-vs-type-1-encryption/

[7] “NSA Type 1 Products vs. Commercial Solutions for Classified (CSfC),” AFCEA International, May 04, 2020. https://www.afcea.org/signal-media/cyber/sponsored-nsa-type-1-products-vs-commercial-solutions-classified-csfc

[8] “Thales CN9120 Network Encryptor,” Thales, Dec. 2024. Available: https://www.thalestct.com/wp-content/uploads/2022/09/cn9120-network-encryptor-PB-12-11-24.pdf

[9] “Commercial Solutions for Classified Program - NSA/CSS,” nsa.gov, 2015. https://web.archive.org/web/20151225183650/https://www.nsa.gov/ia/programs/csfc_program/#dar

[10] “TACLANE E-Series High Speed Encryption Solutions - General Dynamics Mission Systems,” Gdmissionsystems.com, 2025. https://gdmissionsystems.com/products/encryption/taclane-network-encryption/taclane-e-series-high-speed-encryptors

[11] “400G Ethernet Data Encryptor,” Secturion.com, 2024. https://www.secturion.com/products/data-in-transit/darklink-400g-ede/

[12] D. Patel, “Multi-Datacenter Training: OpenAI’s Ambitious Plan To Beat Google’s Infrastructure,” SemiAnalysis, Sep. 04, 2024. https://semianalysis.com/2024/09/04/multi-datacenter-training-openais/

[13] "Optical Transceiver Technology Trends of Data Center in 2022," FiberMall, 2022. [Online]. Available: https://www.fibermall.com/blog/optical-transceiver-technology-trends-of-data-center-in-2022.htm.

[14] A. Douillard et al., “DiLoCo: Distributed Low-Communication Training of Language Models,” arXiv.org, 2023. https://arxiv.org/abs/2311.08105

[15] “Defense-Wide Justification Book Volume 1 of 2,” Defense Information Systems Agency, Feb. 2020. Available: https://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2021/budget_justification/pdfs/02_Procurement/DISA_PB2021.pdf

[16] “NVLink High-Speed GPU Interconnect | NVIDIA Quadro,” Nvidia.com, 2025. https://www.nvidia.com/en-us/products/workstations/nvlink-bridges/

[17] Y. Deng et al., "Characterization of GPU TEE Overheads in Distributed Data Parallel ML Training," arXiv:2501.11771, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2501.11771.

[18] “How does the XOR operation function in the encryption and decryption processes of a stream cipher?,” European IT Certification Academy, Jun. 12, 2024. https://eitca.org/cybersecurity/eitc-is-ccf-classical-cryptography-fundamentals/stream-ciphers/stream-ciphers-random-numbers-and-the-one-time-pad/examination-review-stream-ciphers-random-numbers-and-the-one-time-pad/how-does-the-xor-operation-function-in-the-encryption-and-decryption-processes-of-a-stream-cipher/

[19] “Available TensorFlow Ops,” Google Cloud, 2025. https://cloud.google.com/tpu/docs/tensorflow-ops

[20] “Custom Operators API Reference Guide [Beta] — AWS Neuron Documentation,” Readthedocs-hosted.com, 2023. https://awsdocs-neuron.readthedocs-hosted.com/en/latest/neuron-customops/api-reference-guide/custom-ops-ref-guide.html#custom-ops-api-ref-guide

[21] “NVIDIA Trusted Computing Solutions,” NVIDIA, Mar. 2025. Available: https://docs.nvidia.com/570TRD1-trusted-computing-solutions-release-notes.pdf

[22] “NVIDIA Blackwell Architecture Technical Overview,” NVIDIA, 2024. https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid

[23] “Trainium Architecture — AWS Neuron Documentation,” Readthedocs-hosted.com, 2025. https://awsdocs-neuron.readthedocs-hosted.com/en/latest/general/arch/neuron-hardware/trainium.html

[24] A. Vahdat, “Ironwood: The first Google TPU for the age of inference,” Google, Apr. 09, 2025. https://blog.google/products/google-cloud/ironwood-tpu-age-of-inference/

[25] J. Park, J. Yoo, J. Yu, J. Lee, and J. Song, “A Survey on Air-Gap Attacks: Fundamentals, Transport Means, Attack Scenarios and Challenges,” Sensors, vol. 23, no. 6, p. 3215, Jan. 2023, doi: https://doi.org/10.3390/s23063215.

[26] S. Nevo, D. Lahav, A. Karpur, Y. Bar-On, H. A. Bradley, and J. Alstott, “Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models,” www.rand.org, May 30, 2024. https://www.rand.org/pubs/research_reports/RRA2849-1.html

[27] IBM, “Air gap,” Ibm.com, Oct. 10, 2024. https://www.ibm.com/think/topics/air-gap

[28] “Categorization and Control Selection for National Security Systems,” Committee on National Security Systems, Jun. 2022. Available: https://rmf.org/wp-content/uploads/2022/10/CNSSI_1253_2022.pdf

[29] “Security and Privacy Controls for Information Systems and Organizations | CM-14,” Nist.gov, Nov. 07, 2023. https://csrc.nist.gov/projects/cprt/catalog#/cprt/framework/version/SP_800_53_5_1_1/home?element=CM-14

[30] “Security and Privacy Controls for Information Systems and Organizations | SC-18,” Nist.gov, Nov. 07, 2023. https://csrc.nist.gov/projects/cprt/catalog#/cprt/framework/version/SP_800_53_5_1_1/home?element=SC-18

[31] K. DelBene, M. Medin, R. Murray, and S. Moore, “Zero Trust Architecture (ZTA) Recommendations,” DEFENSE INNOVATION BOARD, Oct. 2019. Available: https://innovation.defense.gov/Portals/63/Templates/Updated%20Meeting%20Documents/ZTA%20Recommendations_CLEARED.pdf

[32] R. Ward and B. Beyer, “BeyondCorp: A New Approach to Enterprise Security | USENIX,” www.usenix.org, Dec. 2014. https://www.usenix.org/publications/login/dec14/ward

[33] D. Kokotajlo, S. Alexander, T. Larsen, E. Lifland, and R. Dean, “AI 2027,” Ai-2027.com, Apr. 03, 2025. https://ai-2027.com/

[34] D. Owen, “Interviewing AI researchers on automation of AI R&D,” Epoch AI, Aug. 27, 2024. https://epoch.ai/blog/interviewing-ai-researchers-on-automation-of-ai-rnd

[35] “Known Exploited Vulnerabilities Catalog | CISA,” Cybersecurity and Infrastructure Security Agency CISA, 2024. https://www.cisa.gov/known-exploited-vulnerabilities-catalog?search_api_fulltext=sandbox

[36] F. Wilhelm, “Project Zero: An EPYC escape: Case-study of a KVM breakout,” Project Zero, Jun. 29, 2021. https://googleprojectzero.blogspot.com/2021/06/an-epyc-escape-case-study-of-kvm.html

[37] B. Adam, “Cross Domain Solutions Overview,” Under Secretary of Defense for Research and Engineering, Feb. 2023. Available: https://www.dau.edu/sites/default/files/Migrate/EventAttachments/838/Slides_DAU%20CDS%20Webinar_Adam_9Feb2023.pdf

[38] “Isode Approach to Data Centric Security using NATO Confidentiality Labels - Isode,” Isode, Jul. 24, 2025. https://www.isode.com/whitepaper/isode-approach-to-data-centric-security-using-nato-confidentiality-labels/

[39] H. Hammer, K. Kongsgard, A. Bai, A. Yazidi, N. Nordbotten, and P. Engelstad, “Automatic Security Classification by Machine Learning for cross-domain Information Exchange,” Military Communications Conference, Oct. 2015, doi: https://doi.org/10.1109/milcom.2015.7357672.

[40] A. Bhatt et al., “Ctrl-Z: Controlling AI Agents via Resampling,” arXiv.org, 2025. https://arxiv.org/abs/2504.10374

[41] R. Greenblatt et al., "AI Control: Improving Safety Despite Intentional Subversion," arXiv:2312.06942, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2312.06942.

[42] Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv:2212.08073, Dec. 2022. [Online]. Available: https://arxiv.org/abs/2212.08073

[43] “Learn About Data Diodes : Owl Cyber Defense,” Owlcyberdefense.com, 2019. https://owlcyberdefense.com/learn-about-data-diodes/

[44] “XML Guard - Isode,” Isode, Jul. 19, 2024. https://www.isode.com/product/xml-guard/

[45] “Activating AI Safety Level 3 Protections,” Anthropic.com, 2025. https://www.anthropic.com/news/activating-asl3-protections

[46] M. Schreiner, “GPT-4 architecture, datasets, costs and more leaked,” *THE DECODER*, Jul. 11, 2023. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

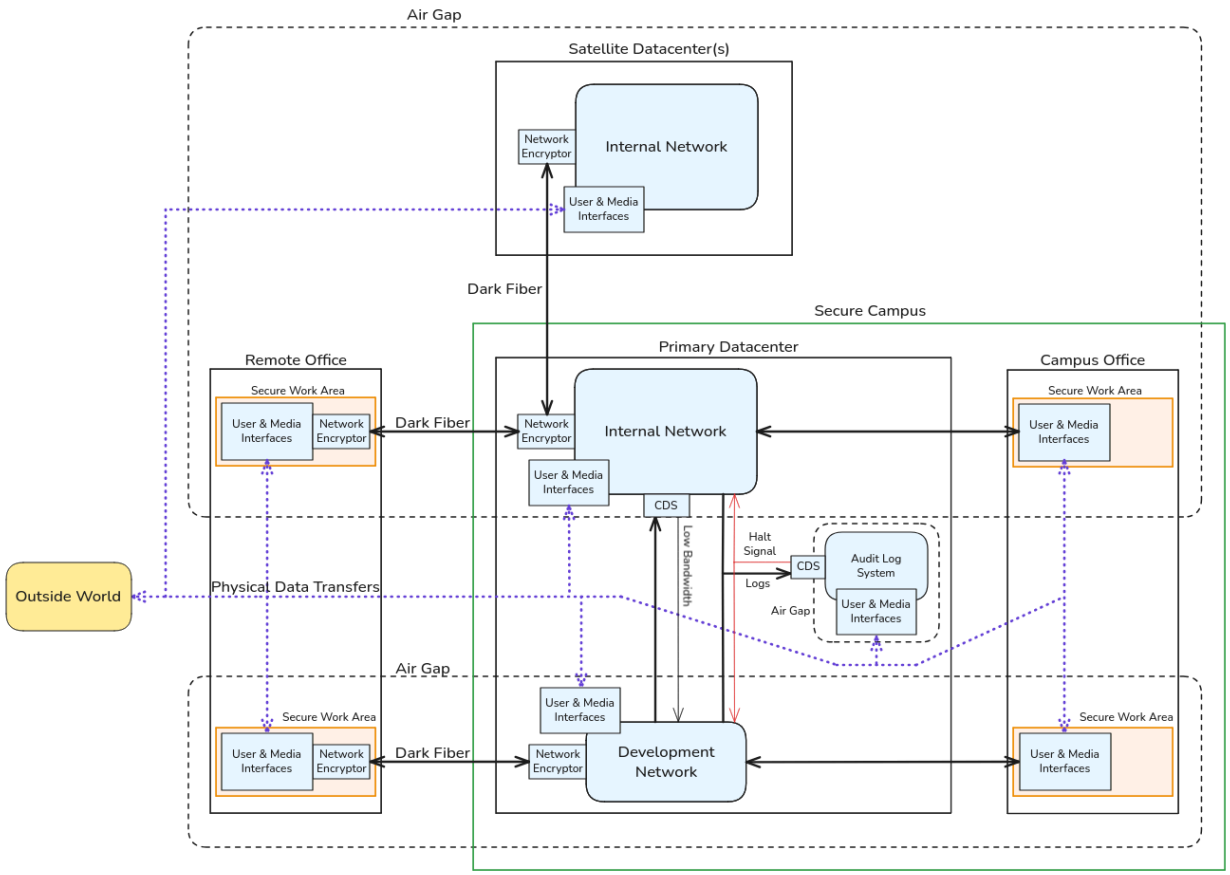
[47] C. Metz, “Google Is 2 Billion Lines of Code—And It’s All in One Place,” Wired, Sep. 16, 2015. https://www.wired.com/2015/09/google-2-billion-lines-codeand-one-place/

[48] J. Harris and E. Harris, “America’s Superintelligence Project,” Gladstone AI, Apr. 2025. Available: https://superintelligence.gladstone.ai/assets/America's%20Superintelligence%20Project%20[Redacted].pdf

Appendix A

These diagrams are not part of the recommendations, but are included to demonstrate what an SL5 network might look like.

Scenario 1: Trusted Long-Distance Connections



Aspects of this diagram not in recommendations:

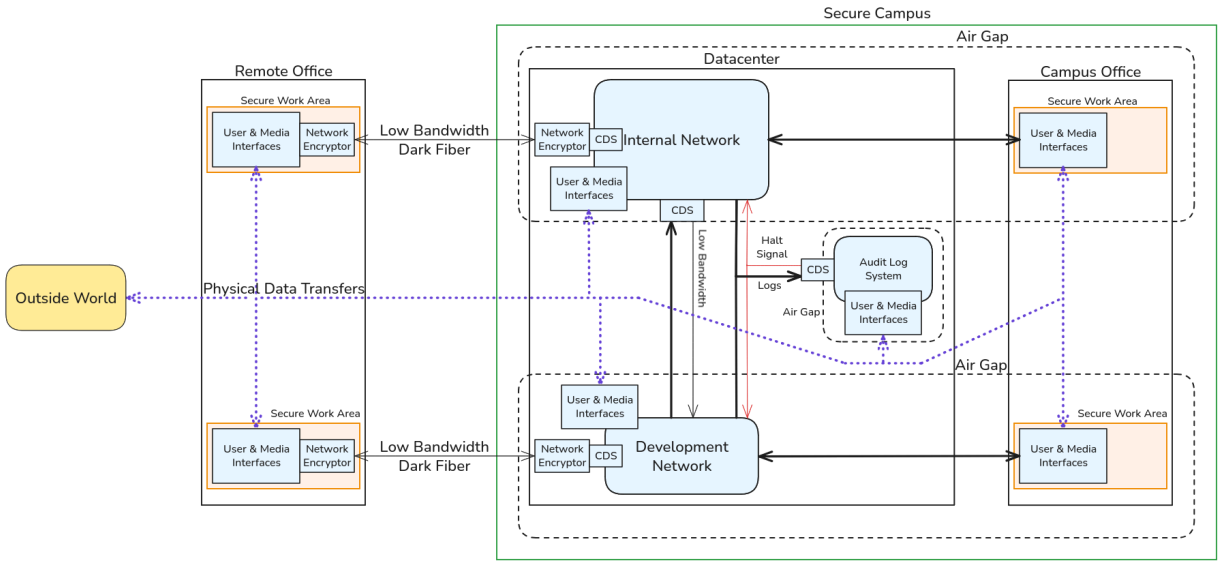
- An audit log system stores logs from both the Development and Internal Networks and has the capability to halt processes in either
- Connections outside physically secured areas are made using dark fiber (note that while useful, dark fiber is vulnerable to highly capable attackers [48])
- Physical transfers allow data to be exchanged with the outside world and within the system

Where do things happen on the network?

- If it requires high bandwidth access to models, it occurs on Internal Network
- If it requires execution of unsigned code, it occurs on the Development Network
- Open question: What workflows might require both, how could these be handled?
- Open question: What workflows might require access to frontier model weights?

- If it requires internet access it occurs on a non-SL5 network, this is out of scope, but parts of office buildings outside the secure work areas could have access to this network to facilitate physical transfer

Scenario 2: No Trusted Long-Distance Connections



Like the previous scenario, except air-gaps are not considered to extend outside physically secured enclaves, and long-distance connections are limited to low bandwidth.

It is also possible to imagine additional campuses connected only by low bandwidth connections and relying otherwise on physical transfers. Some tasks may be highly suited to distribution under such constraints, such as a pretraining run or synthetic data generation.