General Purpose AI Policy Lab

Verification Mechanisms for International Agreement over Frontier AI Development

Lola Carbonell, Pierre Peigné
GPAI Policy Lab · Campus Cyber, Paris

General
Purpose AI
Policy Lab

Part 1

The AI race dynamic is a coordination problem

General Purpose AI Policy Lab

The race to AGI and its consequences

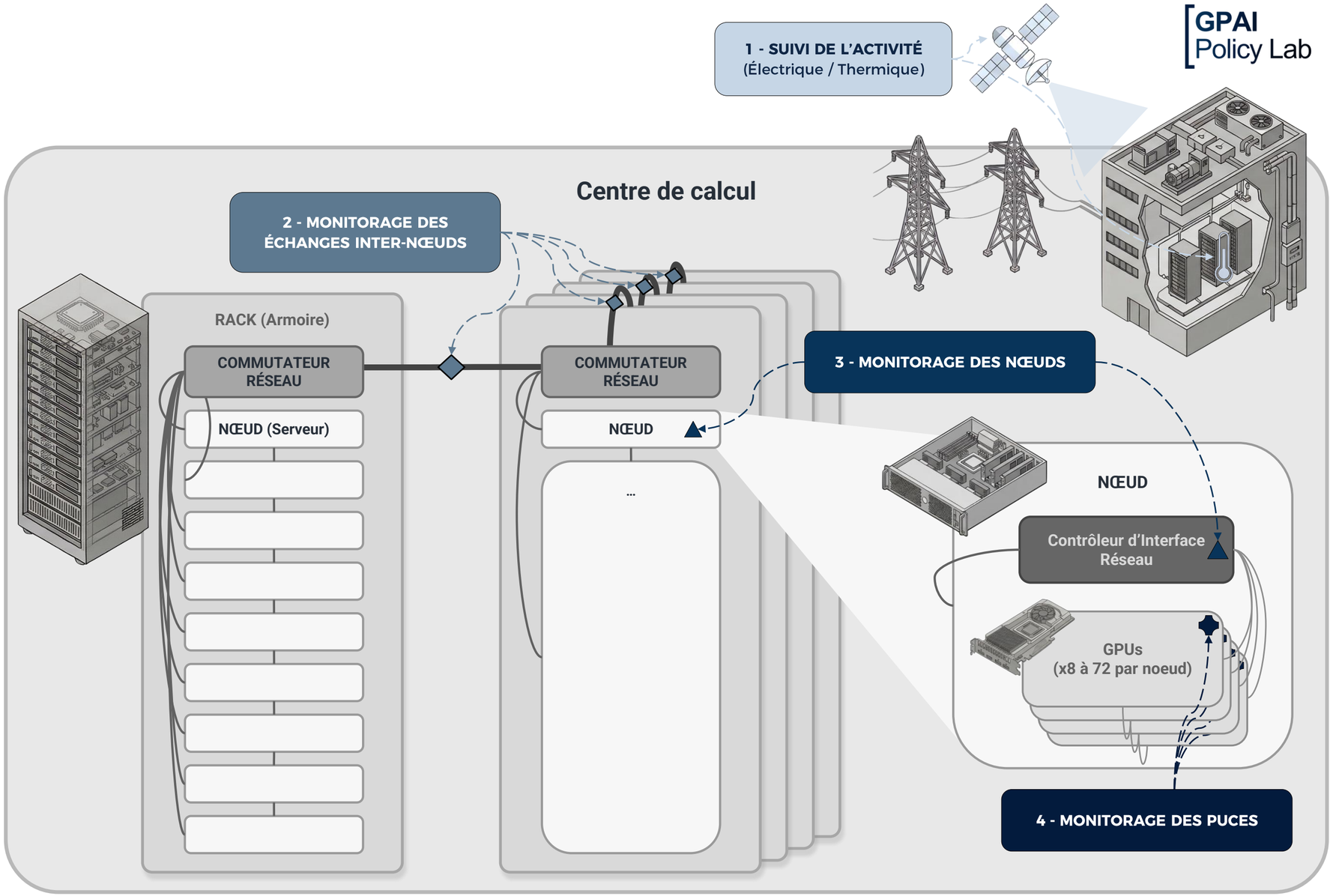
GPAI Policy Lab		China	
		Race to AGI	Differential Development
USA	Race to AGI	→ **Systemic risks** → **Loss of Control**	Perceived US dominance → **Systemic risks** → **Loss of Control**
	Differential Development	Perceived Chinese dominance → **Systemic risks** → **Loss of Control**	**Coordination** → **Control guarantees** → **Benefits from AI**

General
Purpose AI
Policy Lab

Part 2

Overview of the technical solutions for AI development verification

Types of verification mechanisms



©

Timelines for verification mechanisms deployment

GPAI Policy Lab

Suivre l'activité des centres de calcul (puissance électrique, température, etc.)

Avantages
Mise en place quasi-immédiate.

Inconvénient
Mesure très grossière: l'auditeur sait uniquement que le centre de calcul est utilisé, mais pas *pour quoi*.

Vérifier la puissance de calcul par des preuves cryptographiques

Avantages
Mise en place relativement rapide. Preuve que les propriétés déclarées correspondent aux observations.

Inconvénient
Limité à des propriétés simples telles que les FLOPs. Requiert un ancrage physique (network TAP)

Suivre les calculs au niveau des noeuds (par contrôleur d'interface réseau)

Avantages
Visibilité intra-noeud : peut observer l'activité des GPUs d'un noeud de manière individuelle.

Inconvénient
Nécessite de la R&D firmware/hardware au niveau des contrôleurs d'interface réseau. Dépendance aux fabricants (US).

Suivre les calculs au niveau des GPUs

Avantages
Visibilité maximale : peut observer toutes les opérations au sein des GPUs.

Inconvénient
Nécessite de la R&D hardware avancée au niveau des GPUs. Dépendance aux fabricants (US, Chine).

Aujourd'hui

18-36 mois

5 ans

8 ans et plus

©

General
Purpose AI
Policy Lab

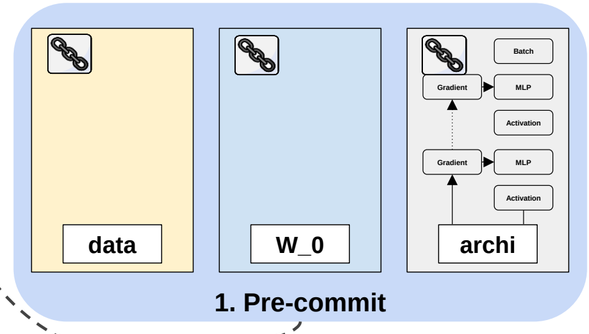
Part 3

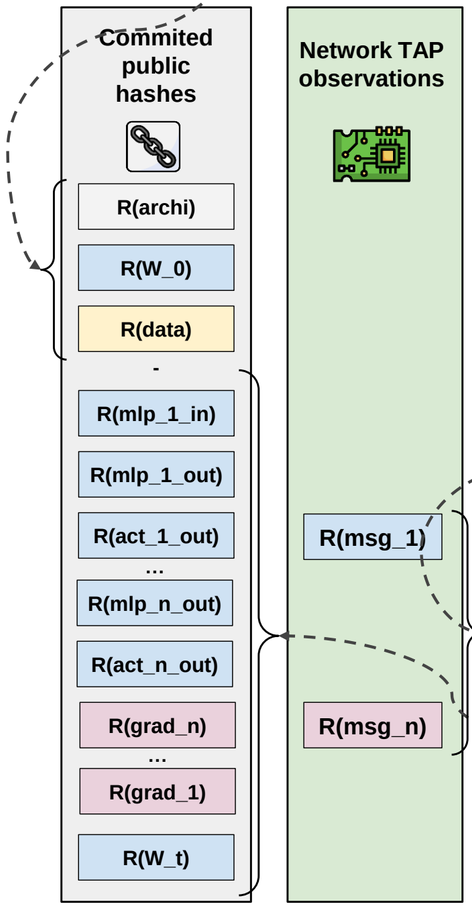
Our approach: sampling-based zero knowledge proofs

General Purpose AI Policy Lab

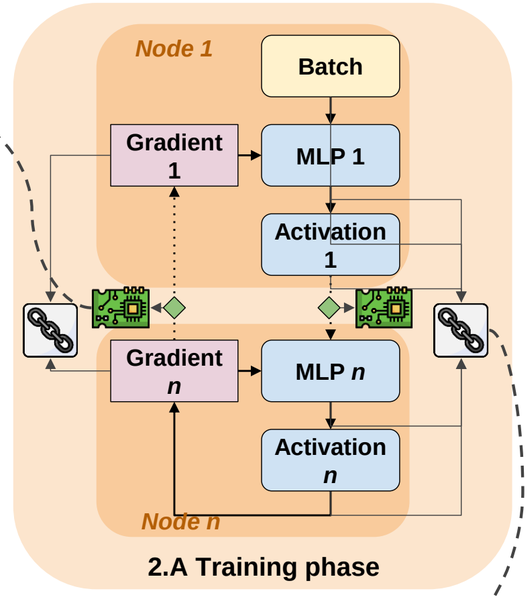
Sampling-based zero-knowledge verification

The trainer pre-commits its dataset, initial weights and architecture *Merkle roots* to a public ledger.

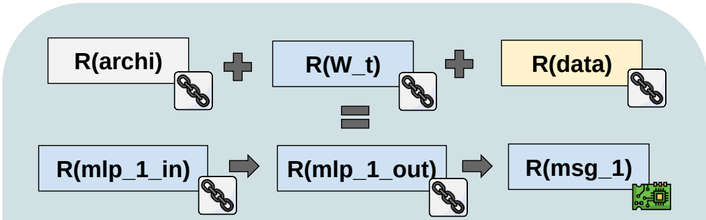




During training, network TAP collect and hash inter-node communications.

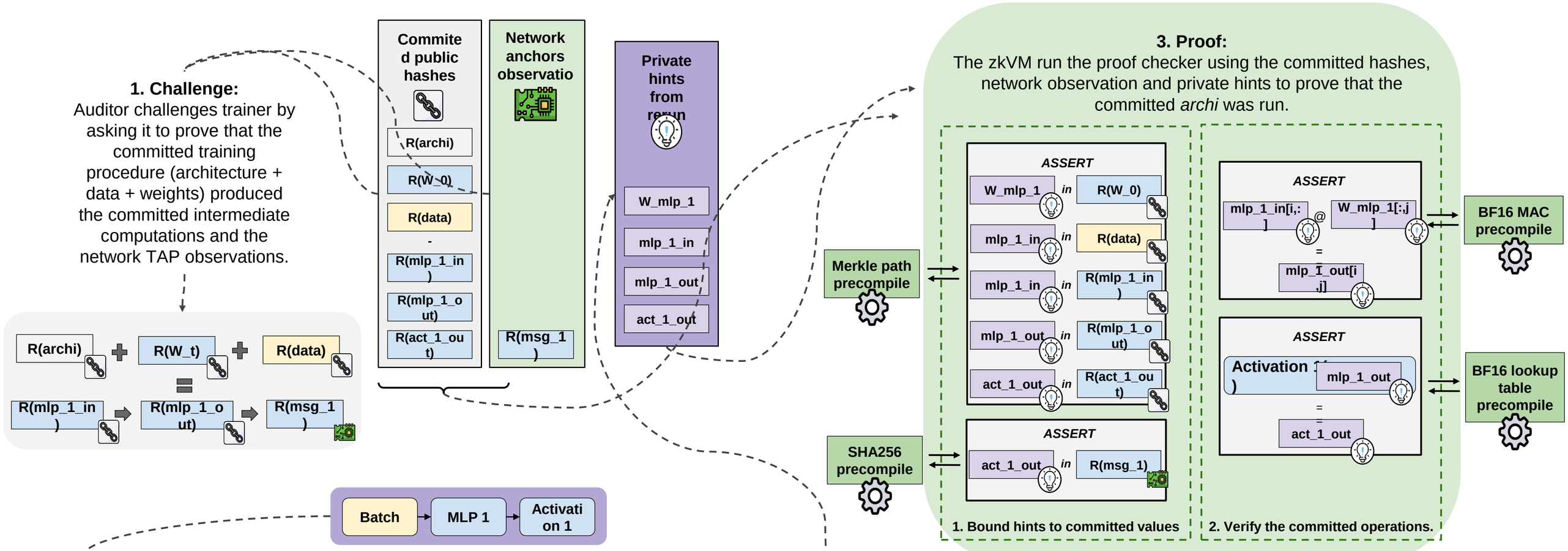


During training, intermediate computations are hashed and committed to the public ledger.



2.B Challenge (during training):
During training, auditor can challenge trainer by asking it to prove that the committed training procedure (architecture + data + weights) produced the observed communications and committed intermediate computations.

Example of implementation with a zkVM



The proving system

How does the proving system work?

- **Genesis proof**
  - Binds the model weights, training procedure and dataset to their precomitted signatures

- ***Ex-ante attestation**:*
  - Binds the model weights, training procedure and dataset to specific training properties we want proof about

- **In-training step proof**
  - Ensures the committed training procedure with proven training properties is the one being run over the cluster

General
Purpose AI
Policy Lab

Questions?