Load-based proxy balancing

The virtual pool sends GPU machines where your employees need them

Explanatory brief for IT management

Cellule-PRO, version 0.2.158

Key points

Your Cellule-PRO servers (the *pools*) are simple servers. The computing power comes from the machines with a graphics card (the *proxies*). Instead of tying each machine to one site forever, you can let it report to the cluster's *virtual pool*, which sends it where people are waiting for answers: first to a pool that has no engine left, then to the busiest pool. When things calm down, it is handed back. **Nothing moves without your approval**: everything is disabled by default.

CELLULE — David MOURGUES · pro.cellule.ai · contact@cellule.ai

Contents

1 In two minutes . . . 2
2 How it works . . . 2
2.1 The three rules of the virtual pool . . . 2
2.2 A proxy's journey . . . 3
2.3 What if a pool goes down? . . . 3
3 The settings, step by step . . . 3
3.1 Step 1 — choose which proxies can move . . . 4
3.2 Step 2 — turn on load-based moving . . . 4
3.3 Step 3 — the switchover delay in case of failure . . . 4
3.4 Step 4 — check in the Supervision . . . 5
4 Advantages and drawbacks . . . 5
4.1 Advantages . . . 5
4.2 Drawbacks and limits . . . 5
5 Should you turn it on? . . . 6
6 Frequently asked questions . . . 6

1 In two minutes

Picture a taxi company with three agencies. Each agency has its own front desk (the *pool*), but the cars (the *proxies*, your machines with a graphics card) do not have to stay parked in front of the same desk all day. In the morning, one agency is swamped while another sits idle; the dispatch radio (the *virtual pool*) then sends a free car to where customers are queuing. In the evening, the car returns to dispatch.

Cellule-PRO does exactly that with computing power:

- **The pools** receive questions from the employees on their site. They are ordinary Linux servers, with no graphics card required.
- **The proxies** are the powerful machines that produce the answers (one or more graphics cards, one or more models).
- **The virtual pool** is the cluster's single address (the virtual IP address, or VIP, held in turn by one of the pools). This is the address the proxies report to.

Key points

- **Only the proxies move.** Never the employees' workstations, never the pools, never the data.
- **No pool is left out.** A pool with no inference engine left is served first.
- **Each pool keeps at least one proxy per model**: we never strip one site to equip another.
- **Two switches, disabled by default**: one for the cluster, one per proxy.

2 How it works

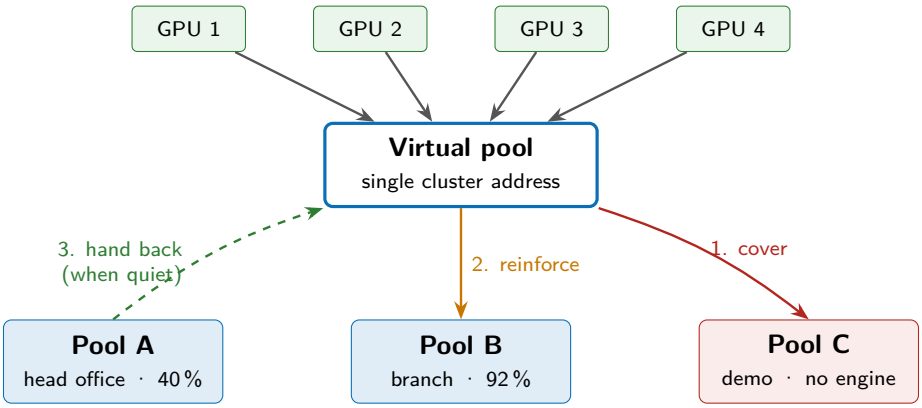
2.1 The three rules of the virtual pool

Every minute, the pool that holds the virtual address looks at the proxies *that arrived through this address* and at the load of each pool in the cluster. It applies three rules, in this order:

Rule	What happens
1. Cover	A pool in the cluster has **no inference engine** (a demonstration pool without a graphics card, a freshly installed pool, a pool whose machines have left): it receives a mobile proxy **first**, even if the virtual pool is itself busy.
2. Reinforce	A pool goes above the **high threshold** (80 % by default): it receives a mobile proxy, provided the virtual pool itself stays below the high threshold without that proxy.
3. Hand back	A pool stays below the **low threshold** (60 % by default) for the minimum time (5 min): it hands back to the virtual pool a proxy it had been lent. Never the last proxy of a model, never the only proxy of a pool.

A pool's load is the share of its inference engines busy answering: 3 busy engines out of 4 = 75 %. A pool with no engine counts as 100 %: it cannot answer anyone.

2.2 A proxy's journey



Employees at each site keep writing to their own pool; only the GPU machines change place.

> **The pool handles it**
>
> A proxy that is told to leave **first finishes the answers in progress** (up to one minute), then reconnects to the destination pool within a few seconds, over an encrypted link. Employees notice nothing except an answer that arrives faster.

2.3 What if a pool goes down?

- **The pool holding the virtual address goes down**: the address moves to another pool within a few seconds (keepalived). The proxies, still running, reconnect to the same address; the new virtual pool places them again. No central database is lost: there is none.
- **Another pool goes down**: the proxies it was hosting switch to another pool in the cluster after the switchover delay (5 min by default, adjustable); they therefore keep answering, and the virtual pool places again those that come back through its address.
- **A pool comes back** with no engine: the "cover" rule sends it a proxy within the next minute.

3 The settings, step by step

Four settings, in this order. The first three are done **only once, from any pool**: they are copied to every pool in the cluster.

First of all: install the proxies on the virtual pool's address

When installing a proxy, enter **the cluster's virtual address** (or its domain name) as the pool address, not the address of one particular pool. A proxy installed on the address of a specific pool stays attached to that pool: it works, but the virtual pool will not place it.

3.1 Step 1 — choose which proxies can move

Tab “Connected workers”, card “Proxies movable according to load”. Tick the “Mobile” column for each machine you agree to move.

Connected workers › Proxies movable according to load

Proxy	Model	Connected	Mobile
`gpu-salle-serveur-1`	35B	to this pool	☑ movable
`gpu-salle-serveur-2`	35B	to this pool	☑ movable
`gpu-vision`	vision	to this pool	☐ stays on its pool

Tip: keep *on its pool* any machine that a site must keep no matter what (for example the only document-reading machine at a site), and make the “backup” machines mobile.

3.2 Step 2 — turn on load-based moving

Tab “Worker balancing”, card “Cluster load and proxy placement”. Tick “Move proxies according to load (cluster)”, check the thresholds, then save.

Worker balancing › Cluster load and proxy placement

☑ **Move proxies according to load (cluster)** enabled (acts only on “mobile” proxies)

High threshold 80 % Low threshold 60 % Minimum time on a pool 300 s

Setting	Default	Effect
High threshold	80 %	Above it, a pool receives reinforcement. Lower = reinforcement sooner, more moves.
Low threshold	60 %	Below it, a pool hands back a reinforcement. Keep a gap of at least 15 to 20 points with the high threshold to avoid back-and-forth moves.
Minimum time on a pool	300 s	How long a proxy stays where it was sent before it can leave again. Longer = more stable.

The table above these settings shows, for each pool, its load, its waiting questions and its engines (busy / total): watch it for a few days before turning the feature on, to choose thresholds that fit your activity.

3.3 Step 3 — the switchover delay in case of failure

Tab “Advanced settings”, category “IT assistant (alerts)”, setting “Delay before a proxy switches to another pool (seconds)”.

Advanced settings › IT assistant (alerts)

Delay before a proxy switches to another pool (seconds) 300 from 30 to 3600

Short (30 to 60 s): fast switchover in case of a real failure, but a network cut of a few seconds may move proxies. Long (5 min or more): more stability. This delay does not apply to proxies installed on the virtual pool’s address, which follow the address within a few seconds.

3.4 Step 4 — check in the Supervision

Open the "Supervision": a machine that has moved appears under its new pool at the next refresh. The history of orders (proxy, origin, destination, status) can be viewed in the "Worker balancing" tab, card "Migration history", and each decision of the virtual pool is written to its log with its reason (*cover*, *reinforce* or *hand back*).

> **Caution**
>
> The installation profile also applies to placement: a proxy is never sent to a pool with a different profile (see the "Installation profiles" brief). A fully air-gapped cluster therefore never lends a machine to a pool that is open to the outside.

4 Advantages and drawbacks

4.1 Advantages

Fewer graphics cards to buy	A machine serves the site that needs it right now, instead of sitting idle at a quiet site. You size your fleet for the cluster's peak, not for the sum of each site's peaks.
No site left without answers	A pool that loses its machines, or a new site without a graphics card, is served first.
Resilience to failures	A pool goes down: its machines keep running and are redistributed. No central server, no single database.
A single address to give out	Each proxy is installed with the cluster's address; adding a site does not require reinstalling any machine.
You stay in control	Two switches disabled by default, a choice made machine by machine, adjustable thresholds, at most 6 moves per hour, a history of every order.
Nothing else moves	No workstations, no data, no accounts: employees always write to their own pool.

4.2 Drawbacks and limits

The network between sites	A machine sent to a remote site answers over your inter-site link (encrypted). On a slow link, each answer arrives a little later than locally; the text of the questions and answers travels over this link.
A few seconds of transition	During a move, the machine finishes its answers and then reconnects: it does not take new questions during this short moment.
A measurement that is still global	Load is measured for the whole pool, not model by model: a pool saturated on document reading but quiet on text may not receive the right reinforcement. Per-model measurement is planned for a future version.

A virtual address is required	Without a virtual address (a single pool, or no keepalived), there is no virtual pool, and therefore no placement.
Nothing happens until it is turned on	Switch off: a pool without an engine stays without an engine. An alert “this pool has no inference engine” is planned in the Supervision.
Machines that change site	For an audit, the machine that answered a question is no longer necessarily the site’s own. The order history lets you trace it.

5 Should you turn it on?

Your situation	Advice
A single pool	Pointless: there is nowhere to send a machine.
Several pools at the same site (same building, fast network)	**Yes.** The benefit is greatest and the network cost is zero.
Several sites connected by a good link	**Yes**, keeping one machine “on its pool” at each site, and making the backup machines mobile.
Sites connected by a slow link	With care: raise the high threshold (85 to 90 %) and lengthen the minimum time, so that machines only move in case of real saturation or a pool without an engine.
Demonstration or evaluation pool	**Yes**: it often has no graphics card, and the “cover” rule lends it a machine.

6 Frequently asked questions

Can a machine leave while its users are in the middle of a conversation?
It finishes the answers in progress before leaving (up to one minute). The conversation itself is kept by the employee’s pool, which does not move.

Can a specific machine be prevented from moving?
Yes: leave its “Mobile” box unticked. That is the default for every new machine.

Is the virtual pool a “master”?
No. It is a temporary role tied to the virtual address: if the pool holding it stops, another one takes it over and carries on, without losing anything, because there is no central state.

What happens if two pools are very busy at the same time?
The virtual pool first serves a pool with no engine at all, then the busiest one, one order at a time, without ever pushing itself above the high threshold and without removing the last proxy of a model.

Are my workstations (employees’ CPUs) moved?
Never. Only the proxies, those dedicated machines with a graphics card, can be moved.