PITON Global

LIVE CHAT SUPPORT OUTSOURCING · PHILIPPINES

The concurrency dividend: the executive economics of live chat outsourcing to the Philippines.

An analysis of the only service channel where one agent can be three, the ceiling where that multiplier breaks, the psychology of the ninety-second reply, and vendor-selection discipline for the channel that sits closest to the buy button.

AUTHORS

John Maczynski — Chief Executive Officer
Ralf Ellspermann — Chief Strategy Officer

JULY 2026
MANILA · ESTABLISHED 2001

Contents

ABOUT THIS SERIES — VOLUME 7

Chat is the channel procurement most consistently misprices — because its unit of capacity is not a seat but a concurrency slot, and the two differ by a factor of three when the floor is run well. This volume prices the multiplier, the ceiling where it breaks, and the revenue that rides on both. The full series lives at piton-global.com.

THE PITON-GLOBAL LEADERSHIP TEAM

About the authors

John Maczynski

Chief Executive Officer

In thirty-plus years approving Fortune-500 CX budgets, John watched chat go from novelty widget to the channel sitting nearest the checkout — and watched vendor after vendor price it like voice with a discount. The concurrency arithmetic in this paper is the arithmetic he made providers defend in

Ralf Ellspermann

Chief Strategy Officer

Ralf built and ran chat floors in the Philippines when concurrency dashboards had to be invented rather than bought. Twenty-five years in-market taught him exactly where the third simultaneous conversation stops being free capacity and starts being three degraded ones — the ceiling this paper

procurement rooms. At PITON-Global he applies it for buyers directly, from first scoping call to certified launch, with no account team in between.

j.maczynski@piton-global.com
+1 402-598-8740

maps. When he reads a provider floor today, the first thing he pulls is the concurrency-versus-quality curve, because he spent a career drawing it.

r.ellspermann@piton-global.com
Manila, Philippines

PITON Global

ABOUT PITON-GLOBAL

Founded in Manila in 2001, PITON-Global advises exclusively for the buyer: no delivery floors of its own, no placement commissions, no supply-side consulting to muddy the recommendation. Its working asset is a live, continuously re-verified map of the Philippine provider market — over 1,000 operations — from which each client receives the six to ten providers genuinely qualified for their program, then a fully transparent, competitive RFP among them. Tripadvisor, TSYS, Garmin, HealthPartners, Yetipay, Workrise, and Norman Disney & Young are among the buyers who have used the practice.

SECTION 01

Executive summary

-61%	**2.8×**	**<45s**	**5.8×**
Fully-loaded cost per productive hour vs. US onshore	Sustainable concurrency on vetted floors — per seat, per hour	First response time held by top-tier chat operations	First-year ROI, reconstructed case (Section 6)

Chat is the only service channel with a built-in multiplier. A voice agent is one conversation; a chat agent, properly supported, is two and a half to three — which means the true unit of chat capacity is not the seat but the concurrency slot, and any vendor quoting chat by the seat is either underpricing their quality or overpricing your program. That multiplier is also the channel's trap: pushed past its ceiling, three conversations become three apologies, and the cheapest channel quietly becomes the one that loses carts at the moment of purchase.

The Philippines runs more outsourced chat than anywhere else, on the written-English depth that makes typed nuance possible at scale. Five findings:

1. **The dividend is real and measurable.** At 2.8 sustained concurrency, a $11.20 Manila chat hour produces resolution capacity that would cost roughly $80 of onshore voice time. The multiplier stacks on the wage delta — Section 2 does the compounding.

2. **The ceiling is where programs die.** Between 2.5 and 3 concurrent conversations, quality metrics hold; past it, per-chat handle time stretches, error rates climb, and CSAT drops 8–14 points. Weak vendors sell concurrency of 4+ as efficiency; the floor read exposes it as queue-shifting onto the customer.

3. **Ninety seconds decides the revenue.** A shopper who opens chat mid-checkout is the highest-intent visitor on the site. Answered under 45 seconds, vetted floors convert 22–34% of pre-sale chats; past three minutes, the cart is usually gone. Chat speed is not a service metric wearing a costume — it is a conversion rate.

4 **Written craft is the hiring bar.** A chat transcript is a permanent, forwardable artifact of your brand. Floors that hire for written-English precision, test for tone under multitasking, and QA the transcript — not just the outcome — hold 90+ CSAT at full concurrency. Floors that staff chat with voice overflow do not.

5 **Bots moved the goalposts, not the humans.** As chatbots absorb the FAQ tier, what reaches a live agent is higher-stakes, higher-emotion, and closer to money. The human layer shrinks in volume and grows in consequence — making agent selection, and therefore vendor selection, worth more per seat than it was five years ago.

The intended reader owns e-commerce, CX, or support operations where chat carries sales-adjacent traffic. Sections 2–4 price the channel; sections 5–7, the method and the proof.

SECTION 02

The concurrency dividend: one agent, three conversations, and where the ceiling sits

Figure 1 plots the relationship every chat contract silently depends on: what happens to quality as an agent's simultaneous conversations increase. The data comes from production floors instrumented during PITON-Global engagements.

FIG. 1 — QUALITY VS. CONCURRENCY: WHERE THE MULTIPLIER BREAKS (VETTED FLOORS, 2025–26)

1–2	2.5–3	3.5–4	4+
Under-loaded	**The dividend zone**	**The stretch**	**The mirage**
CSAT 93+ · capacity wasted — you are buying voice economics in a chat channel	CSAT 90–93 · FRT <45s · per-resolution cost at its floor — where vetted operations run	Handle time +40% per chat · errors climbing · CSAT slides 5–8 points — savings now borrowed, not earned	CSAT −8–14 pts · abandoned chats · repeat contacts — the queue has been moved onto your customer

Source: concurrency instrumentation across PITON-Global chat engagements, 2025–26; CSAT and FRT measured per concurrency band at constant intent mix.

The compounding is what makes chat the best-priced channel in CX when it is run inside the dividend zone. Take the blended Manila chat hour from Section 4 — $11.20 — and divide it by 2.8 sustained conversations: an effective $4.00 per conversation-hour, against $28.90 for a single-threaded onshore voice hour. That is an 86% unit-cost reduction, of which only 61 points come from wages. The remaining quarter is operational — earned by concurrency management, and lost the moment a vendor chases it past the ceiling.

What holds an operation in the zone is machinery, not restraint: intent-aware routing that never stacks three complex cases on one agent, live concurrency dashboards supervisors actually watch, macros that carry judgment rather than replace it, and QA that scores the transcript under load. Ask a prospective vendor for their concurrency-versus-CSAT curve. The good ones have it on one click; the rest quote you a seat price — which tells you they have never measured the thing they are selling.

The ninety-second window: response psychology and the revenue side of chat

A customer on chat has not left the page — they are holding two intentions at once: buy, or leave. Figure 2 traces what first-response time does to the first intention, from checkout-adjacent chats across instrumented client programs.

FIG. 2 — PRE-SALE CHAT CONVERSION BY FIRST-RESPONSE TIME (INDEXED, 2025–26)

Index: purchase completion of checkout-adjacent chats vs. <45s baseline = 100. PITON-Global client programs with e-commerce attribution, 2025–26.

The window reframes what a chat program is for. On vetted floors, 22–34% of answered pre-sale chats convert, average order values run 10–15% higher when an agent guides the choice, and proactive invitations fired on exit-intent recover carts that no retargeting email will ever get back. Priced honestly, a well-run chat desk on a commerce site is not a cost center with a service level — it is a sales channel with a wage bill, and usually the cheapest one the company operates.

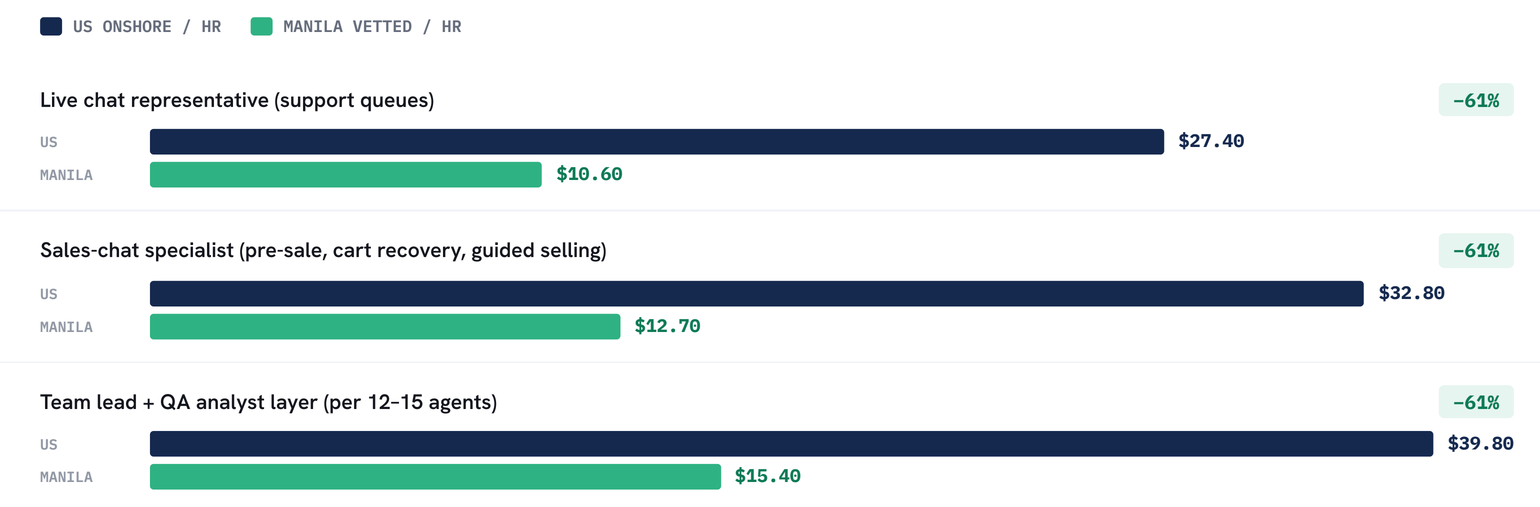
This is also where the concurrency ceiling and the revenue meet: the agent juggling a fourth conversation is precisely the agent whose next reply slips past ninety seconds on a checkout chat. The dividend zone of Section 2 is not a quality preference — it is where the conversion curve of Figure 2 stays on its plateau. Vendors who manage concurrency to protect FRT on revenue intents, and let it stretch only on low-stakes queues, understand the channel; vendors with one global concurrency setting do not.

SECTION 04

Cost and performance: role-band economics and top-tier

benchmarks

TABLE 1 — FULLY-LOADED COST PER PRODUCTIVE HOUR BY ROLE BAND (US$, 2026)



Basis: PITON-Global chat engagement pricing 2025–26 vs. US metro contact-center labor, fully loaded; onshore comparator single-threaded voice-equivalent; 24/7 at flat rates.

TABLE 2 — LIVE CHAT OPERATING BENCHMARKS, VETTED VS. BASELINE (2025–26)

METRIC	VETTED TOP TIER	BASELINE	EXECUTIVE READING
First response time	<45s	2–5 min	The conversion window of Fig. 2, held or lost
Sustained concurrency	2.5–3.0	1.5 or 4+	Under is waste; over is queue-shifting
CSAT at full concurrency	90–93	76–84	Quality measured under load, not at idle
Pre-sale chat conversion	22–34%	8–15%	The revenue side, staffed for deliberately
Chat abandonment in queue	<3%	12–25%	Every abandonment is a customer mid-intent

Source: PITON-Global chat engagement data 2025–26; baseline = typical pre-engagement operations, onshore or unvetted offshore.

Read together: the rate delta gets a program to −61%; the concurrency dividend takes the unit cost to −86%; and the benchmark deltas decide whether the channel sells or leaks. A twelve-point conversion gap on pre-sale chats, priced at any realistic traffic level, is worth more annually than the entire chat contract — which is why Section 5 treats the floor read, not the rate card, as the decision document.

SECTION 05

Sourcing discipline: the 7-Step Vendor Vetting Framework, applied to live chat

Chat diligence hinges on artifacts most vendors have never been asked for: the concurrency-versus-CSAT curve, transcripts read cold, and FRT distributions split by intent. The funnel below is representative of a 20–60-seat chat search.

FIG. 3 — THE 7-STEP FRAMEWORK AS A LIVE-CHAT FUNNEL

	STEP	WHAT IT ELIMINATES IN A LIVE-CHAT SEARCH	FIELD REMAINING
1	Requirements definition	Seat-count thinking: rebuilds the program in concurrency slots by intent — which queues carry revenue, where FRT must be protected, what the bot tier hands off and in what state	1,000+
2	Market mapping	Voice-first floors that treat chat as overflow, and giants for whom a 35-seat written program ranks nowhere	~50
3	Capability screening	Floors without chat machinery: intent-aware routing, live concurrency dashboards, written-English hiring bars, transcript-level QA, sales-chat experience with measured conversion	~14
4	Compliance & security audit	Weak posture for a channel that pastes order numbers and card-adjacent data all day: PCI scope, DLP on transcripts, retention rules, access governance on the chat platform	~10
5	Operational diligence	The chat floor read: pull the concurrency-vs-CSAT curve, read twenty transcripts cold against brand voice, watch the dashboard during a live peak, split FRT by intent, check what happens at handoff from the bot	6–9
6	Competitive RFP & negotiation	Per-seat beauty contests: finalists bid in concurrency slots with FRT floors by intent, CSAT-under-load gates, abandonment ceilings and conversion targets — all carrying credits	2–3
7	Transition & launch governance	Brand-voice drift and concurrency creep at cutover: golden-set transcript certification, concurrency caps enforced in the platform from day one, PITON-Global embedded buyer-side to steady state	1

Field-remaining counts represent a 20–60-seat live-chat search, PITON-Global engagements 2024–26.

The standing guarantees hold: **vendor neutrality** — PITON-Global operates no chat floors and takes no placement economics, so the shortlist reflects curves and transcripts, not inventory; **right-sizing** — providers for whom your program is a flagship account. Six to ten genuinely qualified providers, delivered within 48–72 hours, free to the buyer, competing through a fully transparent, fair, comprehensive, and highly competitive RFP.

SECTION 06 · ANONYMIZED CASE STUDY

From queue-shifting to selling: a 35-seat chat rebuild, reconstructed

Engagement LC-024 is a US online fashion retailer — roughly $240 million in revenue, chat carrying 55% of support volume and most pre-sale questions — whose incumbent offshore vendor had discovered concurrency the wrong way: agents held at 4.5 simultaneous chats, FRT averaging just over three minutes at peak, 19% queue abandonment, and a chat-conversion rate the e-commerce team had stopped reporting out of embarrassment. Identifying details are anonymized under NDA.

SELECTION (WEEKS 0–7, INCL. CHAT FLOOR READS)

Requirements rebuilt the program in concurrency slots across a 19-intent taxonomy: revenue intents capped at 2 concurrent chats with a 30-second FRT floor, service intents at 3 with 60 seconds.

TRANSITION (MONTHS 3–6, CAPS ENFORCED IN PLATFORM)

Concurrency caps were configured into the chat platform itself — not left as policy — from the first routed conversation. Agents certified against a 50-transcript golden set for brand voice and

Mapping produced 48 claimants; the machinery screen — routing, dashboards, transcript QA, measured sales-chat — left 12; the PCI-and-DLP audit left 9. Floor reads, with twenty transcripts read cold at each finalist and the concurrency curve pulled live, shortlisted 6. A Manila floor with fashion-retail chat history won at $11.20 blended, bidding conversion targets with credits attached.

guided-selling craft; proactive exit-intent invitations launched in month 4 on product and cart pages; the bot tier was re-scoped to hand off with context instead of restarting the conversation. PITON-Global chaired weekly buyer-side governance to steady state, then quarterly.

TABLE 3 — LC-024 FIRST-YEAR VALUE BUILD-UP (US$)

VALUE COMPONENT	YEAR 1	BASIS
Labor arbitrage on the 35-seat program	$1.16M	**65.5k productive hrs × ($28.90 − $11.20)**
Pre-sale conversion lift (11% → 26% of answered revenue-intent chats)	$0.38M	Incremental completed orders vs. prior-year cohort, margin basis
Abandonment collapse (19% → 2.6%) — customers answered mid-intent	$0.19M	Recovered chats × historical intent mix, margin basis
Repeat-contact reduction from one-touch resolution at sane concurrency	$0.12M	Avoided contact volume at blended rate, net of ramp
Gross value created	**$1.85M**	
Less: engagement & transition costs	($0.32M)	Golden-set certification, dual-running, platform reconfiguration, governance overlay; advisory fee $0 (provider-paid model)
Net value ÷ cost = first-year ROI	5.8×	$1.85M ÷ $0.32M

Figures rounded; conversion and abandonment effects cohort-verified and taken at margin. Method in Methodology & notes.

Sixty-three percent arbitrage, thirty-seven percent operational value — and note what produced the operational share: *lowering* concurrency from 4.5 to a capped 2–3. The cheaper program was the one that stopped over-loading its agents, because the revenue lines of Table 3 only exist inside the dividend zone. The e-commerce director's line at the year-one review: "We used to think chat was where questions went. It turns out it's where decisions happen."

SECTION 07

The executive playbook: governance for the first 180 days

A chat transition succeeds when the concurrency discipline is enforced by the platform and the contract, not by good intentions — client in command throughout.

DAYS 0–30
Contract the slots, configure the caps

Buy concurrency slots, not seats. Put FRT floors by intent, CSAT-under-load gates, abandonment ceilings and — on commerce queues — conversion targets into the contract with credits. Configure concurrency caps into the chat platform before the first conversation routes; policy-only caps erode within a quarter.

DAYS 30–90

Certify agents on a golden set of transcripts — brand voice, accuracy, guided selling — before volume scales. Turn on proactive invitations only after reactive FRT is stable; a slow floor firing invitations manufactures its own

Certify the transcript, launch the revenue motion

abandonment. Publish the weekly dashboard: FRT by intent, concurrency, CSAT-under-load, conversion.

DAYS 90–180

Tune the bot boundary, certify steady state

Review the bot-to-human handoff monthly: what the bot should absorb, what it should never touch, and whether context survives the transfer. Verify the program through one real peak — a sale event or seasonal surge — before certifying steady state on two clean months. Hand over on documented criteria; keep the quarterly transcript read standing.

THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT

Buying chat by the seat instead of the slot · letting concurrency creep past the ceiling as "efficiency" · one global concurrency setting across revenue and service intents · staffing chat with voice overflow instead of written-craft hires · firing proactive invitations from a floor too slow to honor them. All five are settled at selection and contract.

The live chat argument, in one sentence: this is the one channel where an agent can be three people and where ninety seconds separates a sale from an abandoned cart — a vetted Philippine floor delivers both sides of that equation at an 86% unit-cost advantage, for buyers who read concurrency curves instead of rate cards and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. The dividend is real. We make sure you collect it, not lend it back.

METHODOLOGY & NOTES

How the numbers were built

Concurrency bands (Fig. 1) derive from platform instrumentation across PITON-Global chat engagements, 2025–26: CSAT, FRT and handle time measured per concurrency level at constant intent mix, so the curve isolates load rather than case difficulty. Sustained concurrency counts active conversations per agent-hour, not peak snapshots.

Conversion figures (Fig. 2, Table 2) come from client programs with e-commerce attribution: checkout-adjacent chats joined to order completion, indexed against the under-45-second baseline. Cost figures are fully loaded — wages, benefits and statutory charges, facilities, technology, management and QA overhead, attrition-driven recruiting — net of shrinkage; US comparators reflect metro contact-center labor on the same basis, single-threaded.

The case study (LC-024) is anonymized under NDA; volumes and dates are generalized, figures rounded. Conversion and abandonment effects are verified against prior-year cohorts and taken at margin, not revenue. ROI = net first-year value ÷ all engagement-related client costs. PITON-Global's advisory fee is provider-paid and appears at $0 in the client cost base; neutrality safeguards are documented at piton-global.com.

Market context draws on IBPAP reporting and PITON-Global's proprietary provider mapping (1,000+ operations; chat capabilities re-verified June 2026). Estimates and projections are PITON-Global analysis and should be cited as such. The full series is available at piton-global.com.

What is your chat floor's real concurrency curve?

Thirty minutes with John will tell you. Free to buyers, strictly vendor-neutral — a curve-verified shortlist within two to three business days.

j.maczynski@piton-global.com
+1 402-598-8740