Reference Plant

REFERENCE FAB 2.0 · METHODOLOGY & DEMONSTRATION

Systematic cost & investment target-setting — for a fab that has no reference.

The Reference Fab set cost targets top-down and cut production cost by more than 25%. But it assumed a reference existed. FabONE — the first 300 mm graphene-photonics fab — has none. Reference Fab 2.0 extends the method with an entitlement model, investment targeting and AI.

The Reference Fab — Infineon, **Plieninger, Müller, Ehm & Reczek (2001)**

Live demonstration · **referenceplant.com**

From a proven benchmarking method to the operating model for a first-of-its-kind fab.

THE METHOD

Standardize away geography and asset age so a plant is judged on **how it is run**. The Reference Fab set best-in-class targets and drove sites to catch up — Infineon cut cost **>25%**.

THE BREAK

A first-of-its-kind line has **no reference** to benchmark against. FabONE adds graphene photonics on-chip (BEOL) at 300 mm — a technology with no precedent.

REFERENCE FAB 2.0

Build an **entitlement** from the 200 mm platform, physics and learning curves; add **AI** and **investment targeting**. The catch-up curve becomes the FOAK→NOAK ramp.

Sources: the Reference Fab (Plieninger et al., IEEE/SEMI ASMC 2001); FabONE (Black Semiconductor / Exyte). The method, the evidence and the FabONE application follow.

The industry benchmarks cost-per-unit; but raw cost-per-unit cannot be compared across fabs.

SITUATION

Cost/good-unit is the headline manufacturing metric. Fabs are ranked and funded on it, node by node.

COMPLICATION

Raw cost conflates **manufacturing performance** with three things a plant does not choose day-to-day: **factor-cost geography** (labour, energy), **technology generation**, and **equipment age**. Leachman & Hodges measured 3–5× spread across comparable fabs.

QUESTION

How do you compare two fabs **like-for-like** — isolating how well each is run from where it happens to be?

Source: R. C. Leachman & D. A. Hodges, "Benchmarking Semiconductor Manufacturing," *IEEE Trans. Semiconductor Manufacturing*, 9(2):158–169, 1996.

Hold the physical route constant, re-price it at one reference, and read the gap lever by lever.

STEP 1

Reference route

For each node, a fixed bill of physical activity per wafer across nine process modules — labour-minutes, energy-kWh, equipment and material euros. Process physics; independent of location.

STEP 2

Reference prices

Value that activity at one common factor set — **€35/h** loaded labour, **€0.12/kWh** electricity. Equipment is costed from the node generation and the plant's age; local prices are removed.

STEP 3

Six-lever bridge

Morph the reference plant into the target one attribute at a time. Each step's euros attribute to exactly one lever, and the six deltas sum precisely to the total gap.

Source: Reference Plant engine (`lib/cost.ts`). Quantities synthetic, calibrated to published wafer-cost economics and fab energy intensity (~1.2 kWh/cm²).

Every cost gap decomposes into six levers — two are geography, four are the plant's own performance and assets.

GEOGRAPHY
People cost
Loaded labour rate by site.

GEOGRAPHY
Energy cost
Industrial electricity price by site.

ASSET
Equipment age
Depreciation runs off with tool age.

OPERATIONS
OEE
Utilization absorbs fixed cost over throughput.

OPERATIONS
Overheads
Indirect / facilities cost per wafer.

OPERATIONS
Yield
Good-unit scrap, the quality lever.

Standardization removes the two **geography** levers. The four that remain are what the plant is accountable for — and OEE and yield are kept distinct so utilization and quality never double-count.

Source: Reference Plant six-lever model. OEE here = availability × performance; the quality component is carried by the yield lever.

Dresden looks 18% dearer than Kulim on raw cost — and is the cheaper operation once geography is removed.

RAW € / GOOD WAFER

Raw winner: **Kulim**, by €548 — cheap labour & energy, older written-off tools.

STANDARDIZED € / GOOD WAFER

Standardized winner: **Dresden**, by €527 — better yield, OEE and overheads.

Source: Reference Plant engine. Both fabs 28 nm / 300 mm. Dresden: yield 91%, OEE 74%, age 6 yr. Kulim: yield 82%, OEE 62%, age 9 yr.

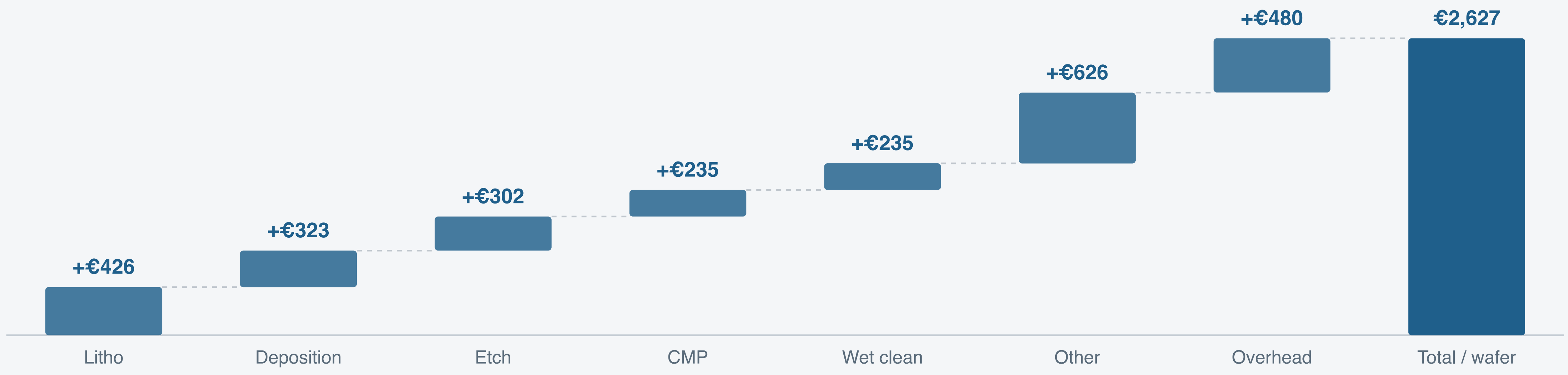
The total bridge shows what moves Dresden off the reference plant — €543 of it is geography that standardization removes.

Source: Reference Plant `referenceBridge()` · 28 nm. Reference plant = reference factor prices, 5-yr tools, 88% yield, 85% OEE. People + energy = €552 location premium (18% of raw cost).

Kulim's €548 raw edge is €1,251 of cheaper geography and assets, masking €702 of weaker operations.

Source: Reference Plant `comparisonBridge()`. The six deltas sum to the €548 gap exactly (verified). Strip the three cheaper levers and the operations levers decide it.

Cost builds process by process — lithography, deposition and etch carry half of a 28 nm wafer before overhead.



Equipment in each module is costed from the node generation and the plant's age, so the module-level bridge isolates **equipment vintage** — concentrated in the equipment-heavy front end. Location is removed, keeping every module comparable like-for-like.

Source: Reference Plant module route · 28 nm at reference factor prices, 5-yr tools. "Other" = ion implant, thermal/diffusion, metrology, sort. Per good wafer = per-wafer ÷ yield.

The nine reference modules, ranked — the numbers behind the process bridge.

Process module · 28 nm	Reference € / wafer
Lithography	€426
Deposition (CVD/PVD/ALD)	€323
Etch	€302
CMP / planarization	€235
Wet clean	€235
Metrology / inspection	€165
Thermal / diffusion	€156
Sort / wafer test	€155
Ion implant	€150

READING THE MODULE BRIDGE

The front-end modules — lithography, deposition, etch — are where equipment cost concentrates, so a plant's **tool vintage** shows up here first. Labour-heavy modules (sort, clean) move with the **people** lever; energy-heavy modules (thermal, etch) with the **energy** lever.

Source: Reference Plant module route, 28 nm, reference factor prices, 5-yr reference tool age. Synthetic, calibrated.

Equipment cost is set by node generation and plant age — an order of magnitude from mature planar to leading-edge EUV.

NEW-PLANT EQUIPMENT COST · € / WAFER

3 nm (GAA, EUV)	€7,500
7 nm (EUV)	€4,700
16/14 nm (FinFET)	€2,150
28 nm (planar)	€1,100
180 nm (power)	€400

DEPRECIATION VS PLANT AGE

Tools depreciate over the asset life (~7 years). A new fab carries the full capital charge; a fully-written-off fab carries only maintenance, ~35% of new.

The honest read on Dresden vs Kulim: Kulim's older tools are *cheaper* per wafer (more depreciated) — a real €229/wafer advantage. Its newer-equipment rival wins not on depreciation but on the operational levers: OEE and yield.

Source: Reference Plant node table & `equipmentAgeFactor()`. Absolute levels synthetic; component shares and node scaling calibrated to public foundry cost estimates.

Labour and electricity are read from the site, not typed in — the geography the method removes.

Representative fab site	Loaded labour · €/h	Industrial electricity · €/kWh
Dresden · Germany	55	0.20
Leixlip · Ireland	45	0.16
Phoenix · United States	48	0.078
Kumamoto · Japan	35	0.17
Singapore	33	0.13
Giheung · South Korea	31	0.11
Hsinchu · Taiwan	27	0.10
Shanghai · China	18	0.09
Kulim · Malaysia	12	0.088

Source: Indicative, calibrated to Eurostat industrial electricity prices (2024) and public manufacturing-wage data. Yardstick inputs, not measured plant data.

Wafer cost rises ~6.5× from 28 nm to 3 nm — but design and mask cost rises ~15×, so fixed cost dominates the leading edge.

Node	Wafer · 300 mm	Mask set	Design / NRE
28 nm	~$3,000	~$2–3 M	$30–50 M
7 nm	~$9,500	$10–15 M	$100–217 M
5 nm	~$18,500	$15–25 M	$200–416 M
3 nm	~$19,500	~$40 M	$400–590 M

MODEL VALIDATION

The synthetic 28 nm reference bill in the engine — **€2,451/wafer ≈ $2,700** — lands right at the public **~$3,000** anchor. Advanced-node totals are shown illustratively; the fixed-cost surge is why chiplets and multi-project wafers exist.

Sources: Wafer: TrendForce / Morgan Stanley (analyst estimates, Jul 2026); mask/NRE: IBS / Handel Jones; SemiAnalysis; CSET. Wafer +~6.5×, fixed cost (design + mask) +~15× across 28→3 nm.

Back-end cost spans four orders of magnitude — from cents for a wire-bond QFN to $1,000+ for a CoWoS-L 2.5D module.

Package technology	Cost / unit	Where it sits
Wire bond	cents–$1	commodity QFN / BGA
Flip-chip (FCBGA)	$1–tens $	mainstream logic
Fan-out (FOWLP / InFO)	$1–tens $	mobile PoP → FOCoS
2.5D (CoWoS)	$300–2,000	HBM / AI · H100 ~$750, B200 ~$1,000
3D (SoIC / Foveros)	+$80–120	stacking add-on

WHY STANDARDIZE BACK-END TOO

Test & assembly is **$80–350+/package** for large AI accelerators — ~$920 (28% of COGS) on an H100. AI demand drove **2025–26 price hikes: TSMC CoWoS +10–20%, ASE +20%**. And **a 1% packaging/test defect cuts margin 5–7%**. Back-end is labour-dominated, so the **people** lever is largest — which is exactly why standardization matters most here.

Sources: Silicon Analysts / Morgan Stanley advanced-packaging cost estimates (Jul 2026); TrendForce (ASE +20%, Jul 2026); Yole / TechInsights. Back-end standardized costing is **live in the demonstration**.

Back-end cost builds the same way — per package, module by module, on the same six-lever bridge.

A flip-chip BGA builds from die attach → bond → mould → ball attach → marking → singulation → final test → burn-in. Back-end is labour- and materials-heavy, so the **people** lever dominates for commodity packages — and the standardized-cost bridge is now **live for assembly & test**, not just wafer fabs.

Source: Reference Plant back-end engine (`lib/backend.ts`) · FCBGA at reference factor prices, 5-yr tools. Per-unit anchors: Yole/TechInsights; IC Insights (2024). Synthetic, calibrated.

Six levers bend the cost curve — each with a live research literature the benchmark can point managers to.

CHIPLET DISAGGREGATION

Split one large die into smaller, higher-yield chiplets; mix nodes. The most-studied fab-cost lever — *Chiplet Actuary* cost model.

YIELD LEARNING (ML)

ML on wafer-map and SEM inspection data cuts scrap and shortens ramp — the largest variable-cost driver.

VIRTUAL METROLOGY & APC

Predict measurements from tool sensors, skip physical metrology steps and catch excursions early.

PREDICTIVE MAINTENANCE (OEE)

Tool downtime spreads fixed cost over fewer wafers; predict failures to schedule maintenance.

TEST & ASSEMBLY VERIFICATION

Front end: multi-site / spatial-sampling test skips redundant coverage. Back end: built-in self-test & on-die impedance sensing verify chiplet/interposer assemblies without external testers.

Alongside cost, the canonical operational KPIs are benchmarked against a peer set, with the correlations an expert expects.

YIELD & DEFECTS

Line yield, die yield, defect density D_0 — die yield from D_0 × area via Murphy's model.

FLOW & EQUIPMENT

Cycle time per mask layer, WIP turns, equipment OEE (SEMI E10/E79).

QUALITY OUT

DPPM shipped, benchmarked against SEMATECH reference ranges.

Each KPI is placed on the peer distribution (p10 / p50 / p90) with direction-aware percentiles, and enforced correlations — low D_0 ↔ high yield, high WIP turns ↔ low cycle time — so the synthetic peer set behaves as a fab engineer expects.

Source: SEMI E10/E58/E79; SEMATECH benchmarks; IEEE IRDS (ex-ITRS); Leachman yield modelling. Peer set synthetic, calibrated to these ranges.

Every figure is synthetic and calibrated to public sources; real customer data would never leave the browser.

SOURCES BEHIND THE MODEL

- **Reference Fab — Plieninger, Müller, Ehm & Reczek (Infineon, IEEE/SEMI ASMC 2001)** — the methodology this demo is built on: ideal → site-specific → catch-up target setting.
- **Leachman & Hodges (1996)** — cross-fab productivity spread, the benchmarking premise.
- **IBS · SemiAnalysis · CSET** — wafer cost by node (28 nm ~$3k → 3 nm ~$19.5k), mask & NRE.
- **Yole / TechInsights; IC Insights (2024)** — advanced-packaging cost tiers; the 1%-defect → 5–7% margin rule.
- **Eurostat (2024); SEMI E10/E79; SEMATECH; IEEE IRDS** — electricity prices, OEE, DPPM, D_0 anchors.
- **Chiplet Actuary (arXiv:2203.12268)** & the cost-lever literature — the model's cost mechanisms.

DATA PROTECTION · REAL DEPLOYMENTS

- **Client-side encryption** — KPIs encrypted in the browser (AES-GCM) before storage.
- **Zero-knowledge benchmarking** — positioned against pre-aggregated distributions; raw values never decrypted server-side.
- **Row-level isolation** — peers are never named; you see distributions, not another company's numbers.

Honesty note: the peer set and per-wafer quantities are generated, not measured, and only calibrated to the ranges above. No proprietary, customer or real-plant data is used in the demonstration.

The method assumed a reference already existed. A first-of-its-kind fab breaks that — so the method evolves.

REFERENCE FAB · 2001

Best-in-class benchmark — a reference exists. Cost split capital · material · personnel · overhead. Annual budget cycle, manual benchmarking. Result: **>25%** cost cut.

REFERENCE FAB 2.0

+ **Entitlement** when no reference exists (physics, equipment limits, lab, learning curves). + **Investment / CapEx targeting**. + **Continuous AI-assisted** loop. + **Agentic data**.

Source: Plieninger, Müller, Ehm & Reczek, "Cost Reduction using Systematic Target Setting of the Reference Fab Methodology," Infineon, IEEE/SEMI ASMC 2001. Extensions = this proposal.

FabONE takes a finished electronic chip and adds a graphene photonic layer on top, at wafer level — and it has no reference fab.

THE TECHNOLOGY

Graphene added **BEOL**, on top of any electronic chip; CMOS-compatible, any node. The **same graphene** forms both modulators and photodetectors.

THE LINE

300 mm (~15,000 m²) built by **Exyte**. Roadmap: pilot line 2026, **pilot production 2027**, volume 2029, full volume 2031.

THE BACKING

€254.4 M raised, **€228.7 M public** via the EU **IPCEI** microelectronics programme. A 200 mm platform (Applied Nanolayers) acquired 2025.

Sources: blacksemi.com; Exyte; EU IPCEI announcements (2024–26). "Integrates high-speed graphene photonics directly onto any electronic chip."

The facility is benchmarkable. The graphene core is not — it needs an entitlement.

BENCHMARKABLE · ~60–70%

Cleanroom, equipment availability, MTBF/MTTR, utilization, throughput, WIP & cycle time, automation, energy, cost per wafer pass. Compared against specialty / BEOL / MEMS / silicon-photonics fabs — not a 3 nm logic mega-fab.

ENTITLEMENT · THE GRAPHENE CORE

Single-crystal graphene, transfer yield, sheet & contact resistance, modulator & photodetector performance, functional optical yield, reliability. No external benchmark — build the entitlement: what is possible once the process is stable.

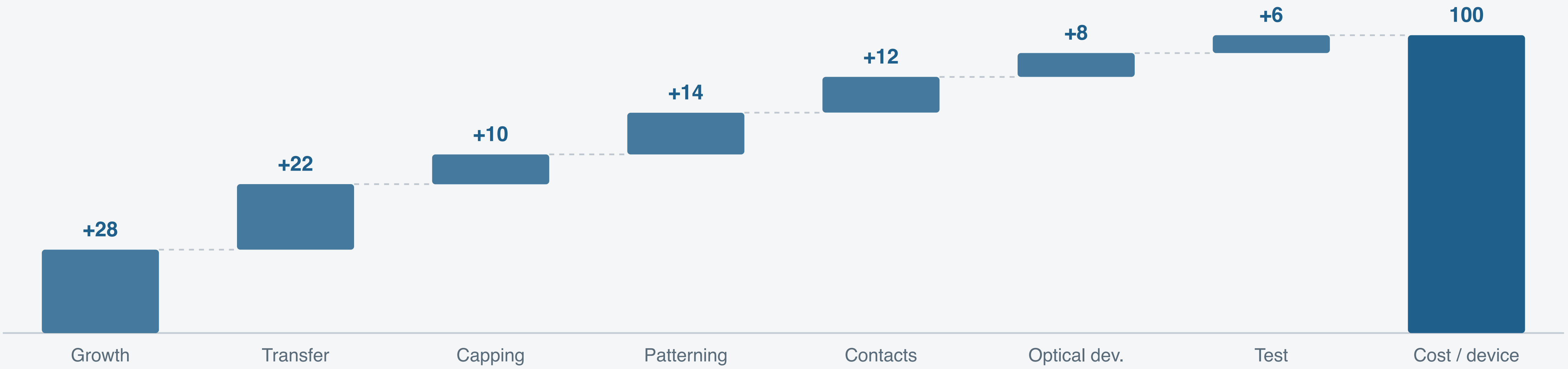
Note: the split of benchmarkable vs entitlement is the honesty of the method — and its value for a technology without precedent.

The entitlement is not a guess — it is normalized to 300 mm from the real 200 mm platform.

Cost per functional device (index, entitlement = 100). The 2027 pilot sits above the entitlement — a gap the ramp closes (the catch-up curve), tracked monthly against plan.

Source: Reference Plant entitlement model · illustrative index, not FabONE data. Anchored to the 200 mm platform (Applied Nanolayers).

Cost builds on the real graphene flow — the steps a conventional fab does not have.



Growth → transfer (BEOL) → capping → patterning (hardmask) → contacts (W damascene) → optical device (modulator + photodetector) → test. Transfer and contacts carry the variability.

Source: follows the published 300 mm CMOS flow (arXiv:2304.02646). Illustrative index (total = 100).

The targets are anchored to what has already been demonstrated at 300 mm.

300 MM, TODAY

Wafer-scale graphene modulators in a 300 mm CMOS platform: **50 dB/mm**, **15 GHz**, across **400 devices**.

YIELD ANCHOR

~**74%** device yield at wafer scale; **>95%** monolayer uniformity on 150/300 mm.

HEADROOM

Intrinsic **200 GHz** RC-limited — the ceiling the entitlement reaches for.

DIFFERENTIATOR

One graphene layer → **modulator + photodetector** (57 mA/W, 64%).

Sources (arXiv): 2304.02646 (300 mm CMOS modulators, 400 devices); 1402.7127 (multifunctional mod+PD); 1302.3854 (broadband PD); 2007.00955 (200 GHz ceiling).

Reference Fab 2.0 is not invented from scratch — each part rests on established literature.

TARGET SETTING & BENCHMARKING

The Reference Fab itself (Plieninger, Müller, Ehm & Reczek, IEEE/SEMI ASMC 2001); cross-fab productivity spread (Leachman & Hodges, IEEE Trans. Semiconductor Mfg 1996).

YIELD LEARNING CURVES

The catch-up curve *is* a yield-learning curve: Tirkel, "Yield Learning Curve Models" (IEEE TSM 2013); Weber, "Yield Learning and the Sources of Profitability" (IEEE TSM 2004).

FIRST-OF-A-KIND ECONOMICS

FOAK→NOAK cost learning: $C_{NOAK} = C_{FOAK}(1–x)^{d}$; FOAK plants run **15–55%** dearer (NETL costing methodology; power-law technology learning).

ML FOR YIELD & COST

Explainable AutoML for yield (arXiv:2403.12381); predictive quality with uncertainty (arXiv:2605.07752); virtual metrology & defect root-cause (referenceplant.com/insights, 18 papers).

Sources: IEEE/SEMI ASMC 2001; IEEE Trans. Semiconductor Manufacturing (1996, 2004, 2013); NETL FOAK/NOAK costing; arXiv 2403.12381, 2605.07752.

A neural network predicts the achievable floor from the process levers — and the investment to get there.

ENTITLEMENT FROM SPARSE DATA

Derive the 300 mm target from limited 200 mm runs + physics — few-shot / transfer learning — instead of waiting for volume.

NEURAL-NET RAMP FORECAST

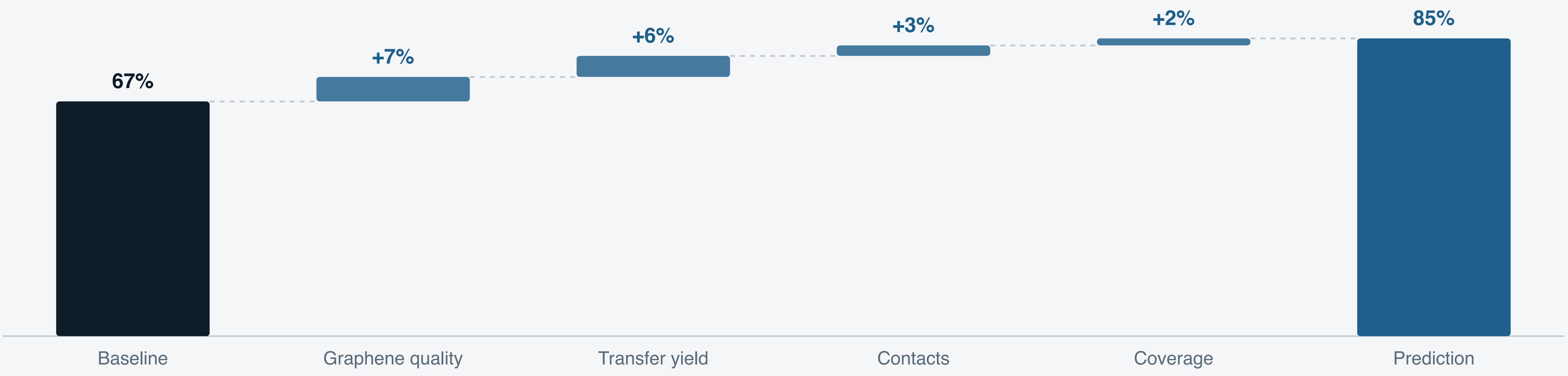
An MLP predicts the 2031 cost floor from the levers; the ramp converges to it. Anomaly detection turns the monthly loop live.

INVESTMENT & CASH-BURN

Move a lever → cost per functional device, tool-sets, CapEx and cash-burn all update together. Phase CapEx to the entitlement ramp.

Live & interactive: the model — levers, neural-net floor, ramp and investment — runs in the browser at **referenceplant.com/entitlement**.

The model shows its work — every prediction breaks down into what each parameter contributed.



Exact **SHAP** attribution: from a neutral **67%** reference, graphene quality (+7), transfer yield (+6), contacts (+3) and coverage (+2) build the **85%** prediction. Not a black box — the network also exposes its sensitivity, partial dependence, an **ensemble confidence band**, and a **cost-aware optimizer** that finds the cheapest way to a yield target.

Live & interactive: the full simulator — SHAP, uncertainty, live network view and optimizer — runs in your browser at **referenceplant.com/simulator**. Uncertainty via ensemble disagreement (arXiv:2605.07752).

A European consortium: the operating model for a first-of-its-kind fab.

THE GAP A CONSORTIUM FILLS

Capital and cleanroom are funded and under construction. What is still missing is a proven way to **set targets, model cost and time investment** for a technology with no precedent.

GRAVITNOMAD'S CONTRIBUTION

Reference Fab 2.0 as FabONE's operating model; the **AI layer** (sparse-data entitlement, ramp forecasting, agentic benchmarking); a live controlling loop process → yield → cost → cash-burn.

Proposal, not a claim: a collaboration model within a STEP/IPCEI-eligible consortium. No affiliation or agreement is implied.

The method is live, bilingual and interactive — describe a fab and read its bridge in under a minute.

ACCESS

referenceplant.com

Sign-in · **demo@referenceplant.com**
Passphrase · **reference2026**

English & Deutsch — one toggle.

WHAT TO TRY

- Open **Dresden Fab 1** — read the six-lever bridge and the process-module table.
- Open **Compare fabs** — put Dresden against Kulim; watch the ranking invert.
- Add a fab — pick a **location** (labour & energy auto-fill) and a **node generation**.
- Generate the **narrative report** — a grounded read-out, in your language.

Note: the demonstration account is pre-loaded with two contrasting 28 nm fabs so the flip is visible on first sign-in.

From a live demonstration to the target-setting system of FabONE — in a European consortium.

NOW

Live & interactive

The method, the entitlement bridge and the neural-net cost model run today at referenceplant.com — move a lever, see the cost, the floor and the investment.

NEXT

Co-develop the entitlement

Calibrate the entitlement on the real 200 mm platform data (Applied Nanolayers); train the forecast on live ramp data.

THEN

FabONE's operating model

Deploy Reference Fab 2.0 as FabONE's cost & investment target-setting system, co-developed with the fab team in a STEP/IPCEI consortium.