GROWNET — Performance & Scalability Plan

(SLO-driven, Cost-Aware, Growth-Ready)

---

1. Objectives & Non-Goals

Objectives

- تضمین تجربه کاربر در رشد
- پیش‌بینی هزینه زیرساخت
- جلوگیری از سقوط سیستم در پیک‌ها

Non-Goals

- Over-engineeringزودهنگام
- بهینه‌سازی برای سناریوهای غیرواقعی

تمرکز **Scalable enough, not perfect** :

---

2. Key Workloads & Traffic Profile

Core Workloads

Workload	Description
User interactions	Dashboard, analytics
Data ingestion	Campaign data
Background jobs	Ranking, sync
API calls	Integrations

Traffic Assumptions (Explicit)

Metric	Current	12-mo Target
DAU	500	25,000
API req/sec	20	800
Concurrent users	50	2,000

---

3. Service Level Indicators (SLI)

Core SLIs

Component	SLI
API latency	p95 response time
Availability	% uptime
Error rate	5xx ratio
Throughput	req/sec

---

4. Service Level Objectives (SLO)

SLI	Target
API latency p95	< 300 ms
Availability	99.9%
Error rate	< 0.5%
Background job delay	< 5 min

هر SLO مستقیماً به **تصمیم فنی یا هزینه‌ای** وصل است

---

5. Capacity Planning Model

Capacity Formula (Example)

Required Capacity = Peak Load × Safety Factor (1.5)

Example

- Peak API load: 800 req/sec
- Required capacity: 1,200 req/sec

Scaling Assumptions

- Stateless services
- Horizontal scaling first
- Vertical scaling only for DB

---

6. Load Testing Strategy

Test Types

Test	Purpose
Baseline	Current load
Stress	Breakpoint
Soak	Memory leaks
Spike	Traffic surges

Tools

- k6 / Locust
- Synthetic traffic mirroring

Frequency

- Before major releases
- Before growth campaigns

---

7. Known Bottlenecks & Mitigations

Bottleneck 1: Database Reads

Risk: Latency under high read load
Mitigation:

- Read replicas
- Query optimization
- Caching layer

Bottleneck 2: Background Processing

Risk: Queue backlog
Mitigation:

- Worker autoscaling
- Job prioritization

Bottleneck 3: External APIs

Risk: Third-party rate limits
Mitigation:

- Circuit breakers
- Async processing

- Backoff strategies

---

8. Scalability Strategy by Layer

Layer	Strategy
Frontend	CDN + static caching
API	Stateless autoscaling
Data	Sharding-ready schema
Jobs	Queue-based scaling

هر لایه **independent scaling** دارد

---

9. Cost vs Performance Tradeoffs

Cost Drivers

- Compute autoscaling
- Database replicas
- Cache size
- Observability tooling

Example Cost Curve

DAU	Infra Cost / month
1k	€400
10k	€1,200
50k	€4,500

رشد خطی، نه انفجاری

---

10. Observability & Alerting

Metrics

- Latency histograms
- Queue depth
- CPU/memory usage

Alerts

Trigger	Action
p95 latency > SLO	Scale up
Error rate spike	Rollback
Queue backlog	Add workers

---

11. Failure & Degradation Strategy

Graceful Degradation

- Disable non-critical features
- Serve cached data
- Read-only mode

اول **بقا**، بعد **کمال**

---

12. Scalability Roadmap

Phase	Focus
Now	Baseline SLOs
+6 months	Autoscaling tuning
+12 months	DB sharding
Scale	Multi-region (if needed)

---

13. Executive Takeaway (Investor Lens)

ما دقیقاً می‌دانیم رشد کجا فشار می‌آورد،
چگونه کنترلش کنیم،
و هر هزار کاربر جدید چقدر هزینه دارد.

این یعنی:

- رشد قابل پیش‌بینی
- ریسک فنی کنترل‌شده
- burn rateقابل دفاع