PROGETTO GIANT: FRAMEWORK BASATO SU DIGITAL TWIN E AI PER LA GESTIONE DINAMICA DELLE RISORSE DEI DATA CENTER

Giovanni Marangi*, Virginia Spinozza **, Alessandro Cosimo Buscicchio **

* Key4 srl – Società di consulenza tecnico scientifica - Contrada Baione, Monopoli (BA), Italy; marangi@key-4.com
** WPS srl – Società di consulenza specializzata in sviluppo software, soluzioni IoT e Industria 4.0, - Contrada Baione, Monopoli (BA), Italy

Il settore dell'*High Performance Computing* (HPC) sta guidando l'innovazione tecnologica e scientifica della nostra epoca. La necessità di avere sistemi di calcolo sempre più potenti e l'avvento dell'era dei supercomputer Exascale hanno evidenziato come l'obiettivo di crescita non possa più essere scisso dalle implicazioni energetiche, operative e ambientali che ne derivano. In un'ottica di ottimizzazione dell'efficienza energetica e della sostenibilità ambientale dei data center moderni, il *framework* GIANT propone una soluzione innovativa basata sulle potenti tecnologie di Intelligenza Artificiale e Digital Twin, che realizza un sistema intelligente in grado di permettere una gestione dinamica delle risorse di questi sistemi tanto potenti quanto complessi da amministrare. L'integrazione con tecniche di *Power Capping* dinamico permette uno *shift* paradigmatico da approccio reattivo ad approccio proattivo ai problemi di efficienza energetica, fornendo uno strumento potente, modulare e scalabile per la gestione dei sistemi HPC moderni.

SUPERCOMPUTER E LA CRESCENTE DOMANDA DI RISORSE COMPUTAZIONALI NELL'ERA DIGITALE

I supercomputer rappresentano l'apice dell'evoluzione tecnologica nel campo del calcolo ad alte prestazioni (*High Performance Computing*, HPC). Questi sistemi sono caratterizzati da architetture complesse composte da centinaia e in alcuni casi anche migliaia di nodi computazionali che lavorano in parallelo e che permettono l'esecuzione di simulazioni e calcoli che sarebbero altrimenti impossibili o richiederebbero tempi proibitivi con sistemi convenzionali. Nell'epoca contemporanea la domanda di risorse computazionali ha seguito un trend crescente esponenziale, alimentato da molteplici fattori che hanno trasformato radicalmente il panorama tecnologico e scientifico globale. La digitalizzazione dei settori produttivi, l'Internet of Things (IoT) e l'avvento di potenti tecnologie data-driven come l'intelligenza artificiale e il *deep learning* hanno imposto la necessità di sistemi capaci di processare, analizzare ed estrarre valore da enormi quantità di dati in tempi sempre più brevi. Questa necessità è trasversale ai più svariati settori dell'industria e della ricerca scientifica, spaziando dalla genomica alle previsioni metereologiche, dall'astronomia alla ricerca

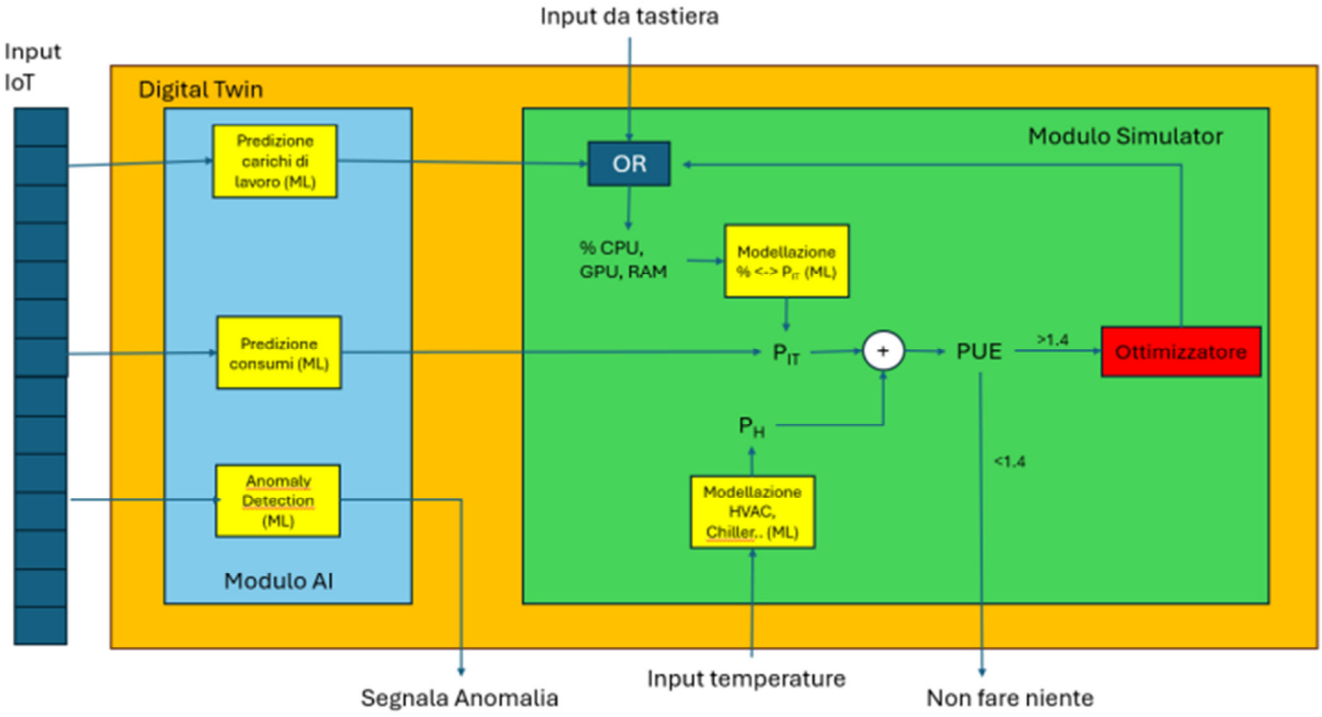

Figura 1 – Il *framework* GIANT

farmaceutica. L'IoT e l'*edge computing* hanno ulteriormente amplificato questa domanda, creando ecosistemi distribuiti dove miliardi di dispositivi connessi generano flussi continui di dati che richiedono elaborazione in tempo reale. I supercomputer diventano così la spina dorsale computazionale dell'industria 4.0 e 5.0 rappresentando uno strumento tanto potente quanto complesso da gestire.

TRANSIZIONE ALL'ERA EXASCALE: PRESTAZIONI RECORD E IMPLICAZIONI ENERGETICHE

Il 2022 ha segnato una pietra miliare storica nell'evoluzione del settore HPC, con l'entrata in produzione di Frontier [1], il primo supercomputer exascale al mondo, sviluppato presso l'Oak Ridge National Laboratory negli Stati Uniti. Frontier è stato il primo sistema in grado di raggiungere prestazioni superiori a 1.1 exaflops, ovvero oltre un quintilione (〖10〗^18) di operazioni in virgola mobile per secondo. Questo traguardo rappresenta l'ingresso in una nuova era computazionale che promette di rivoluzionare la ricerca scientifica e l'innovazione tecnologica. L'incremento di complessità e di potenza computazionale tuttavia porta con sé implicazioni importanti dal punto di vista dei consumi energetici, dell'impatto ambientale e delle emissioni dei data center. La transizione verso l'era exascale ha evidenziato come l'aumento delle prestazioni computazionali non possa più essere perseguito senza considerare attentamente le implicazioni energetiche, operative e ambientali che ne derivano. Quasi tutta l'energia consumata dai sistemi HPC è convertita in calore, quindi in aggiunta all'energia strettamente necessaria per alimentare i nodi computazionali bisogna tenere in considerazione anche i consumi dei sistemi esterni, quali condizionatori, impianti di ventilazione, impianti di illuminazione e pompe dell'acqua [2]. I consumi di questi sistemi ausiliari per l'illuminazione e il raffreddamento del data center rappresentano un fattore di limitazione per le performance energetiche globali del sistema e non possono essere disaccoppiati in quanto sono necessari per il corretto funzionamento del data center.

IL FRAMEWORK GIANT: ARCHITETTURA INTELLIGENTE PER LA GESTIONE DINAMICA DEI DATA CENTER

È in questo contesto che si inserisce il progetto "GIANT: *framework* basato su *digital twin* e AI per la gestione dinamica delle risorse dei data center". GIANT si pone l'obiettivo di realizzare un innovativo sistema intelligente per la gestione delle risorse di questi sistemi complessi ed eterogenei, focalizzandosi sulla massimizzazione dell'efficienza energetica del data center e sulla minimizzazione delle emissioni e dello spreco di risorse energetiche. Grazie alla potenza predittiva dell'intelligenza artificiale e alle capacità di modellazione della tecnologia *digital twin*, il *framework* punta a diventare una soluzione integrata capace di orchestrare dinamicamente l'allocazione delle risorse computa-

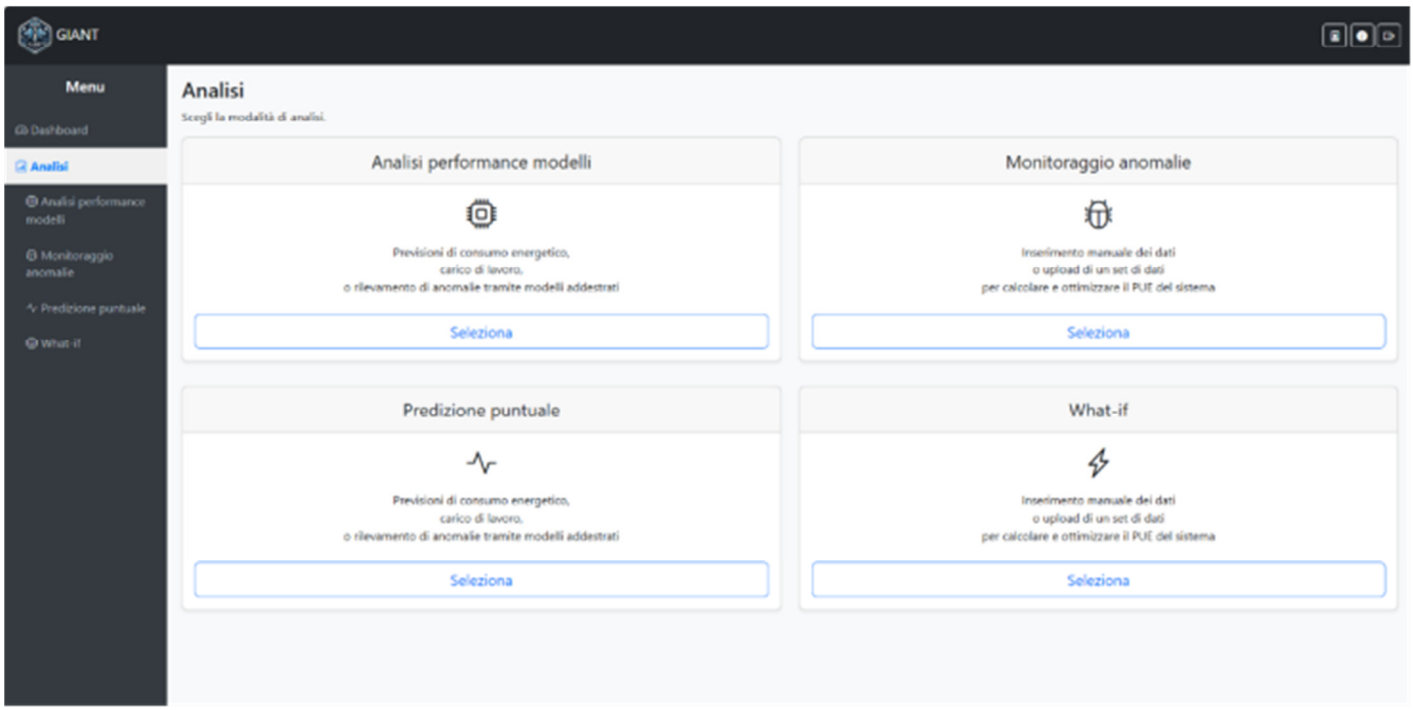

Figura 2 – *Dashboard* della piattaforma GIANT

zionali in risposta alle variazioni dei carichi di lavoro e alle condizioni operative del data center. Sul piano tecnico, l'architettura del *framework* GIANT è strutturata secondo un approccio *multi-layer* che separa le responsabilità funzionali garantendo al contempo una stretta integrazione tra i diversi componenti del sistema. Questa organizzazione modulare facilita la manutenibilità, l'estensibilità e la scalabilità della soluzione proposta, permettendo l'adattamento a diverse tipologie di *data center* e configurazioni infrastrutturali.

ARCHITETTURA *MULTI-LAYER* DEL *FRAMEWORK* GIANT: DALL'ACQUISIZIONE DATI ALL'INTERFACCIA UTENTE

Il primo componente fondamentale è il Data Acquisition Layer, responsabile della raccolta continua di dati dall'infrastruttura fisica del *data center*. Questo strato sfrutta molteplici sensori che monitorano parametri critici di diverso tipo: dalle metriche di utilizzo dei singoli processor, ai consumi energetici fino a parametri ambientali. Il data *acquisition layer* rappresenta le fondamenta del progetto, in quanto raccoglie dati di vitale importanza che permettono l'utilizzo delle potenti tecnologie *data driven* utilizzate nelle fasi avanzate del progetto.

Il secondo componente è il *Cloud Layer*, che svolge un ruolo fondamentale nella gestione e nell'elaborazione dei dati generati dai dispositivi IoT. Nel *cloud*, questi dati sono ricevuti, elaborati, archiviati e resi disponibili per le analisi predittive e correttive e l'accesso da parte degli utenti. Questo *layer* rappresenta il ponte di comunicazione con i due moduli tecnologici principali del sistema: il modulo AI e il modulo Simulator. I dati elaborati sono analizzati dagli algoritmi di *machine learning* basati su XGBoost [3] e reti neurali LSTM (*Long Short-Term Memory*) [4], i quali sono in grado di riconoscere pattern complessi e relazioni non lineari all'interno dei dati di fondamentale importanza per ottenere predizioni accurate. Questi algoritmi permettono al sistema di effettuare *forecasting* dei consumi energetici, dei carichi di lavoro delle componenti del *data center* e di rilevare l'insorgere di anomalie che possono anticipare imminenti guasti e interruzione dei servizi del sistema fisico. Il modulo Simulator rappresenta l'implementazione del *digital twin*, che contestualizza i dati delle previsioni per l'ottimizzare la configurazione delle risorse.

Non essendo stato possibile modellare il comportamento attraverso l'utilizzo di un modello fisico sufficientemente accurato, si è optato per un'implementazione del modulo Simulator per mezzo di due modelli *data driven*. I due sistemi implementati modellano il comportamento dal punto di vista energetico rispettivamente delle componenti IT del *data center* e dei sistemi ausiliari quali impianti di raffreddamento, condizionatori, sistema di illuminazione e pompe dell'acqua. Le informazioni fornite dal modulo Simulator sono di vitale importanza per il corretto funzionamento di GIANT, che grazie all'utilizzo congiunto di queste informazioni con algoritmi di ottimizzazione avanzati come *Sequential Least Squares Programming* (SLSQP) [5] è in grado di individuare la configurazione di risorse ottimale per rientrare nei vincoli imposti in termini di efficienza energetica massimizzando la produttività del sistema riuscendo quindi ad ottenere un duplice obiettivo.

L'ultimo strato è l'*Application Layer*, la componente che consente a tutti gli attori, come utenti o operatori, di interagire con il sistema GIANT. L'interfaccia utente intuitiva permette agli utenti di effettuare analisi delle prestazioni dei modelli di *machine learning* su *dataset* di test e di avere la possibilità di valutarne le performance attraverso la fruizione di grafici dinamici. Il sistema mette a disposizione uno strumento di diagnostica per la sezione di rilevamento delle anomalie, dove oltre ai grafici degli errori di ricostruzione è possibile indagare quale *feature* è risultata essere più problematica per effettuare diagnostiche di sistema e individuare le single componenti critiche che hanno mostrato comportamenti anomali. L'altra funzionalità fondamentale messa a disposizione nell'interfaccia è rappresentata dalla possibilità di effettuare le previsioni sui consumi e sui carichi di lavoro utilizzando il modulo AI. Le previsioni dei modelli di *machine learning* alimentano il modulo Simulator, il quale è in grado di fornire una rappresentazione accurata dello stato futuro del sistema, fornendo informazioni chiave circa l'efficienza energetica futura del sistema. Qualora l'andamento dovesse avere un trend che mira ad andare oltre ai vincoli desiderati, può entrare in gioco il modulo di ottimizzazione, che individuerà automaticamente la configurazione ottimale delle risorse e la suggerirà all'operatore. È questo il vero punto di forza del sistema GIANT: l'interazione tra modelli predittivi e moduli di simulazione permette di implementare un approccio proattivo invece che reattivo, riducendo i tempi di intervento da parte degli operatori e anticipando inefficienze e disservizi, rendendo possibile l'intervento correttivo prima ancora del verificarsi del problema. Il sistema permette anche l'esplorazione di scenari "*what if*", che rendono possibile l'analisi degli ef-

fetti di modifiche alle configurazioni del sistema senza il bisogno di renderle operative sul sistema fisico, riducendo il rischio di guasti e disservizi indesiderati e impattando direttamente sui costi e l'efficienza del data center.

APPROCCIO PROATTIVO E POWER CAPPING: OTTIMIZZAZIONE ENERGETICA E SOSTENIBILITÀ AMBIENTALE

L'adozione di un approccio proattivo nella gestione dei *data center* rappresenta un cambio di paradigma fondamentale rispetto alle metodologie tradizionali di tipo reattivo, offrendo vantaggi significativi in termini di efficienza operativa, riduzione dei costi e minimizzazione dell'impatto ambientale. Mentre i sistemi reattivi intervengono solamente dopo il verificarsi di problematiche o inefficienze, l'approccio proattivo implementato dal *framework* GIANT consente di anticipare e prevenire situazioni critiche attraverso l'analisi predittiva e la simulazione efficiente. La gestione proattiva elimina i tempi di latenza tipici degli approcci reattivi, dove il rilevamento del problema, l'analisi delle cause e l'implementazione delle contromisure possono richiedere intervalli temporali significativi durante i quali il sistema opera in condizioni subottimali. Il *framework* GIANT, attraverso i suoi modelli predittivi, è in grado di identificare trend negativi e situazioni potenzialmente critiche con anticipo sufficiente per implementare azioni correttive prima che si manifestino effetti negativi sulle prestazioni o sull'efficienza energetica. Quando i modelli identificano deviazioni dai pattern normali o trend che potrebbero portare a situazioni problematiche, il sistema può allertare automaticamente gli operatori con sufficiente anticipo per permettere interventi mirati. Il risultato è un sistema che si adatta dinamicamente alle condizioni operative variabili, mantenendo costantemente un equilibrio ottimale tra prestazioni ed efficienza energetica.

La suddetta efficienza energetica è conseguita mediante l'adozione di tecniche di *power capping* intelligente che consiste nella limitazione dinamica della potenza massima consumabile dai componenti hardware per mantenere il consumo totale del sistema entro limiti predefiniti. Tradizionalmente, il *power capping* viene implementato attraverso meccanismi reattivi che intervengono quando il consumo energetico supera soglie predefinite, riducendo la frequenza operativa dei processori o disattivando temporaneamente componenti non critici. Tuttavia, questo approccio reattivo può causare degradazioni improvvise delle prestazioni e instabilità operative, particolarmente problematiche per applicazioni HPC che richiedono prestazioni consistenti e prevedibili. Inoltre il grado di complessità nell'applicazione di queste tecniche aumenta esponenzialmente quando si prendono in considerazioni *data center* molto grandi come I sistemi Exascale. Il *framework* GIANT rivoluziona l'approccio al *power capping*, permettendone

Figura 3 – *Analisi delle prestazioni sulla piattaforma GIANT*

l'applicazione anticipate e graduale.

Dal punto di vista ambientale, l'aumento dell'efficienza energetica si traduce direttamente in una diminuzione delle emissioni di CO_2 per carico di lavoro, contribuendo agli obiettivi di sostenibilità ambientale sempre più stringenti nel settore HPC. Il *framework* GIANT permette ai *data center* di operare con *footprint* carboniche ridotte senza sacrificare le capacità computazionali, supportando la transizione verso un'informatica ad alte prestazioni più sostenibile e dimostrando come l'integrazione di tecnologie di *digital twin* e intelligenza artificiale possa trasformare radicalmente la gestione energetica dei *data center* dell'era exascale.

CONCLUSIONI

Il *framework* GIANT rappresenta una risposta innovativa e necessaria alle sfide crescenti dell'era exascale, dove la complessità gestionale e i consumi energetici dei *data center* hanno raggiunto livelli che richiedono approcci radicalmente nuovi. L'integrazione sinergica di tecnologie di *digital twin* e intelligenza artificiale all'interno di un'architettura *multi-layer* modulare e scalabile offre una soluzione concreta per trasformare la gestione tradizionale dei *data center* da reattiva a proattiva. Attraverso l'implementazione di strategie di *power capping* intelligente e ottimizzazione dinamica delle risorse, il sistema dimostra come sia possibile conseguire simultaneamente obiettivi apparentemente contrastanti: ottenere la massima prestazione possibile minimizzando l'impatto energetico e ambientale. La capacità del *framework* di anticipare criticità operative e di ottimizzare continuamente l'allocazione delle risorse fa di GIANT una tecnologia ottima per la sostenibilità del settore HPC. In un contesto dove la domanda di risorse computazionali continua a crescere esponenzialmente mentre aumentano le necessità di riduzione delle emissioni di CO_2, soluzioni come GIANT non rappresentano più un'opzione, ma una necessità imprescindibile per garantire la sostenibilità economica e ambientale dei *data center* del futuro. L'approccio modulare e la compatibilità con le infrastrutture esistenti facilitano l'adozione della soluzione, aprendo la strada a una nuova generazione di *data center* intelligenti, efficienti e sostenibili che potranno supportare le crescenti esigenze computazionali dell'era digitale senza compromettere gli obiettivi di sostenibilità ambientale globale.

GIANT

Il progetto è stato selezionato e finanziato nell'ambito del Programma di Ricerca e Innovazione dell'Ecosistema "iNEST – Interconnected Nord-Est Innovation Ecosystem" (codice ECS00000043), con il contributo dell'Unione europea – NextGenerationEU, nell'ambito del Piano Nazionale di Ripresa e Resilienza (PNRR), attraverso il bando a cascata promosso dalla SISSA in qualità di Spoke 9, finalizzato allo sviluppo dell'Area Mezzogiorno (codice progetto: 2D1C2C7ED7).

RIFERIMENTI BIBLIOGRAFICI

[1] S. Atchley et al., *Frontier: Exploring Exascale*, in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, in SC '23. New York, NY, USA: Association for Computing Machinery, nov. 2023, pp. 1–16. doi: 10.1145/3581784.3607089.

[2] A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, e L. Benini, *Predictive Modeling for Job Power Consumption in HPC Systems*, in High Performance Computing, vol. 9697, J. M. Kunkel, P. Balaji, e J. Dongarra, A c. di, in Lecture Notes in Computer Science, vol. 9697. , Cham: Springer International Publishing, 2016, pp. 181–199. doi: 10.1007/978-3-319-41321-1_10.

[3] T. Chen e C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, ago. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[4] S. Hochreiter e J. Schmidhuber, *Long Short-Term Memory*, Neural Computation, vol. 9, fasc. 8, pp. 1735–1780, nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[5] J. Nocedal e S. J. Wright, A c. di, *Sequential Quadratic Programming*, in Numerical Optimization, New York, NY: Springer, 1999, pp. 526–573. doi: 10.1007/0-387-22742-3_18.