COSA f2f meeting 18 Maggio 2015

WP3 - Stato Cluster CNAF (A.Ferraro)

Andrea Ferraro presenta il cluster del cnaf, al momento composto da 6 jetson, più varie altre piattaforme (12 in tutto) tra cui 2 schede HMP (Heterogeneus multiplocessing, tutti core della big.little attivi) - si veda tabella in slide 3.
Il cluster usa un server avoton low power come macchina di supporto per NAT, TFTP, DHCP, NFS, lDAP
Acquisito uno switch giga ethernet e un alimentatore con doppia tensione 5/12 + un multimetro da banco e un alimentatore da banco.
Vedi slide per dettagli.

In programma l'installazione di un batch system (lsf o slurm).

Discussione su stato attuale del mercato SoC e possibili nuove acquisizioni. Per i test LHC sarebbe opportuno spostarsi su piattaforme 64 bit - alternative possibili (annunciate) ma non ancora acquisibili nvidia X1 dragonboard410, HIkEY96:
https://www.96boards.org/products/hikey/
https://developer.qualcomm.com/mobile-development/development-devices/dragonboard/410c
http://www.nvidia.com/object/tegra-x1-processor.html

Non è chiaro se per settembre (da quando secondo il proposal dovremmo iniziare a costruire il cluster del cnaf sarà possibile acquisirle). In ogni caso viene deciso di aspettare e non fare ulteriori acquisti a 32 bit. Se necessario chiedere a csn5 di spostare il finanziamento al 2016.

Viene deciso di organizzare una phone conference per discutere le richieste (preventivi) per il 2016. Il portale va compilato a Luglio, quindi preparare un doodle per la settimana del 24 giugno. Daniele dovrà circolare quanto presentato a settembre prima della phone conference.

WP2: Stato test e benchmark

L. Morganti presenta una serie di attività di testing fatte al cnaf e pd. Vengono presentati i risultati con test sintetici e con applicazioni reali su tutte le piattaforme acquisite al cnaf. Con test sintetici la nvidia tegra K1 risulta la migliore come performance per la parte cpu e la più semplice da usare per quanto riguarda l'uso della GPU (grazie alla possibilità di programmare in CUDA). Viene presentato il porting su questa piattaforma di una vasta serie di applicazioni dalla ricostruzione di immagini di xray tomography, alle simulazioni astrofisiche e cosmologiche alla dinamica molecolare. Le performance vengono confrontata con quanto ottenibile su piattaforme tradizionali di classe server in termini di performance assolute e di rapporto performance su watt o su joule. A seconda dell'applicazione la tegra è più lenta da tre a 10 volte ma i consumi sono talmente ridotti da renderla molto competitiva nel rapporto performance su watt – vedi slide per dettagli. L'applicazione di ricostruzione tomografica è particolarmente promettente.

Vengono presentati i risultati per il test di HEPSPEC06 su varie piattaforme arm e su un avoton (low power intel) - la jetson e l'odroid sono particolarmente performanti (meglio dell'avoton) per un solo core caricato.

Discussione su opportunità di confrontare oggetti tipo development board e oggetti di classe server.

Discussione anche sul fatto che le correnti per i server vengono prese a monte dell'alimentatore mentre quelle delle dev board a valle. Si decide di investigare la possibilità di prenderle a valle anche per i server come fatto in alcuni lavori in letteratura misurando sui cavi di alimentazione della motherboard, pci-e e ausiliari verso la gpu.

Vengono discusse alcuni possibili customizzazioni di frequenza e attivazione di cpu (core), gpu e memoria che verranno approfondite nel successivo talk di Enrico Calore.

WP5: Applicazioni e Ambiente Software

Due presentazioni: Tommaso Boccali (LHC) ed Enrico Calore (Lattice Boltzamann).

LHC (Boccali)

Boccali riporta delle esperienze su architetture low power in ambito LHC, in particolare con codici di CMS e ATLAS.

Viene introdotto il test ParFullCMS basato su GEANT4, molto utile per CMS. – I risultati presentati con questo test sono a favore di architetture intel, ma la cpu low power è un xgene-1 – viene quindi deciso di provarlo anche sul cluster del cosa a 32bit (azione su Tommaso di circolare i dettagli di come possa essere girato su arm). Vengono presentati i test con sw PYTHIA, ROOT e un codice di calcolo teorico fatto in collaborazione con univ pisa. K1 più lenta di un fattore 3 rispetto ad intel xeon e circa equivalente ad un avoton – risultati sorprendentemente buoni.

Viene discussa la necessità di avere architetture a 64bit per poter sfruttare il ROOT-IO al momento impossibile su 32 bit ma necessario per superare la fase di test ed usare questi oggetti in produzione

Si discute la possibilità di avere del codice su gcc5.1 per sfruttare arm+gpu

Vengono presentati i risultati con un codice di lattice QCD - la jetson K1 performa oltre le aspettative soprattutto in considerazione della limitata memory bandwidth.

Si discute dell'affidabilità di queste schede low power per lunghi run (settimane) – al momento nessuno ha riscontrato problemi di surriscaldamento per run lunghi 1,2 giorni. La ventola sulla jetson sembra più che sufficiente e si discute della possibilità di rimpiazzarla con un dissipatore passivo.

Lattice Boltzmann (Calore)

E.Calore presente le sperimentazioni fatte con il codice L.B. sviluppato a ferrara e portato su una serie di piattaforme e linguaggi di programmazione: c++(vect 512 bit per mic), c++ (vect 256bit per xeon), cuda, openmp, opencl etc. vedi slide per dettagli.

Viene mostrato un sistema di acquisizione della corrente consumata basato su arduino con possibilità di triggerare l'acquisizione per specifiche funzioni del codice, in modo da misurare il consumo di diverse parti del codice.

Vengono effettuate una lunga serie di prove sulla jetson variando parametri quali la frequenza della cpu, della gpu e della memoria. Vengono costruiti plot 4d alla ricerca del minimo che caratterizza il miglior rapporto prestazioni-consumo energetico. Vedi slide per dettagli.

Discussione sul fatto se nella ricerca del minimo vada considerato il consumo di baseline della schede. Tutti concordi sul fatto che date le conclusioni sia di ferrara che del cnaf debba essere considerato.

Per il codice cuda vengono effettuati anche test al variare di alcuni paramentri sw quali ad esempio la block size, anche in questo caso ricercando il miglior punto di lavoro per quanto riguarda il rapporto performance /watt.

Slide 25 è un ottimo riassunto delle possibilità di customizzazione delle performance della jetson k1.

WP5: Simulazione di reti neurali DPSNN distribuita su due Jetson

Pier Paolucci presente il lavoro di porting ed i risultati ottenuti su jetson vs xeon con un codice di simulazione di reti neurali (DPSNN). La prima parte descrive il codice ed i risultati ottenuti su architetture tradizionali all'interno dei progetti euretile e corticonic. La seconda parte descrive i risultati ottenuti in cosa (vd slide per dettagli) – numeri conclusivi compatibili con quanto presentato nei talk precedenti: 2.2 micro-Joule per simulated synaptic event on the "embedded dual socket node", 4.4 times better than spent by "server platform", instantaneous power consumption: "embedded "14.4 times better than "server" "server" platform 3.3 faster than "embedded".

Discussione su necessità/opportunità di provare l'applicazione su un numero maggiore di nodi, non necessariamente k1, sul cluster del cnaf. Vista la facilità con cui è possibile eseguire i run viene deciso di provare l'applicazione al cnaf. Contatti offline per decidere le modalità.

Discussione sulla possibilità di avere maggiore memoria sulle schede low power e proposta di utilizzare le memorie eMMC presenti su alcune schede come l'odroid che offrono prestazioni paragonabili alla ram. Prova che potrà essere eseguita al cnaf.

Discussione sulla possibilità di migliorare le performance ottimizzando la generazione di numeri casuali che potrebbe essere demandata alla gpu sulle jetson (future work).

Discussione sulla ncecessità di raccogliere i codice su un unico repository. Daniele propone un gitlab privato installato al cnaf ed offerto come servizio nazionale. Autenticazione basata su aai INFN. Circolare i dettagli alla lista COSA.

Viene anche deciso di raccogliere sul sito i report e i risultati pubblicabili per ciascun test.

WP4: Stato cluster e attività ROMA (P.Vicini, A.Lonardo)

Piero Vicini presenta il lavoro svolto a roma dal gruppo APENET+ su interconnessioni basate su FPGA con RDMA.

FPGA SoC(System on Chip) e' un componente ibrido che integra una sezione hardware programmabile e configurabile dall'utente (FPGA) e core(s) di processori lowpower, high performance. Q

Due motivazioni principali per l'adozione di FPGA SoC:

- Accelerazione di task computazionali eseguiti dal uPembedded nella FPGA correlati al protocollo RDMA implementato nell'architettura di rete custom APEnet (APEnet V5, NaNet e derivati, sistemi picoLO)
- FPGA Embedded hardened multiple ARM cores (anche a 64 bit) come processore di calcolo ("accettabile" in termini di prestazioni) caratterizzato da basso consumo di potenza e integrazione direttacon la network. Esplorazione di nuove architetture dedicate (es. DPSNN) ExaNeSt e derivati

(vedi slide per dettagli)

Il cluster basato su fpga della serie ARRIA10 non può ancora essere acquisito a causa del ritardo nella disponibilità dell'hq prevista per q4 2015. A causa del rapporto sfavorevole eur/$ sarà probabilmente possibile acquisire su fondi cosa solo 2 schede invece che 4 ma è possibile economia di scala con il progetto nanet – da discutere nella phone di giugno.

In attesa delle nuove FPGA con ARM cores embedded prosegue lo sviluppo e l'ottimizzazione della versione V5 di APEnet+. In particolare:

- Finalizzazione dell'interfaccia con PCIe gen3 x8 realizzata in due flavors distinti
- PLDA PCIe core (black box, high cost…)
- Altera core nativo: richiesto sviluppo e ottimizzazione del backend necessario in prospettiva per il controllo completo della tecnologia di interfaccia con l'host
- Ottimizzazione dell'interfaccia con la nuova architettura del link seriale
- Ottimizzazione del motore RDMA GPU/CPU
- Dual TX RDMA
- Sviluppo driver in userspace

Vicini presenta il progetto H2020 exanest (approvato) per lo sviluppo di una architettura di un supercalcolatore alla exascale per quanto riguarda la parte di storage e interconnessione a bassa latenza (vedi slide)

A.Lonardo presenta il progetto H2020 picoLO (sottomesso, responso atteso per settembre) – sviluppo di un sistema embedded con prestazioni computazioni elevate per alcune tipi di applicazioni, industriali e non (incluse alcuni applicazioni di interesse INFN, in particolare trigger hw di esperimenti hep). Data la call con forte impronta industriale, il consorzio è equamente distribuito tra partner accademici e industriali.

Vedi slide per dettagli.

Le idee tecniche alla base di picolo sono però estremamente utili per il progetto COSA per quanto riguarda l'interconnessione delle schede low power a bassa latenza per calcolo parallelo hpc con performance pure comparabili con un cluster tradizionale. In particolare la parte di switch pci-e è molto promettente come ulteriore sviluppo R&D di COSA. L'hw che renderebbe possibile questo sviluppo, switch fabric pci-e, è ora disponibile (disponibilità da maggio 2015) con un numero di linee adeguato. Il costo è elevato, ma affrontabile, si fanno proposte per coprire il costo, es. riusare parte dei fondi cnaf qualora non fosse possibile acquisire hw 64 bit o fare un passaggio in commissione5.

Chiaramente se picolo venisse approvato potrebbe ricadere anche sul progetto ma i tempi per le richieste fondi alla commissione non sono compatibili. Punto su cui ragionare e discutere alla phone di giugno su preventivi 2016.

Viene ribadita le necessità, anche in sinergia con i progetti presentati, di acquisire quanto prima (appena disponibili) dev board basate su tegra X1 a 6t4bit) – vedi slide conclusiva del talk Vicini/Lonardo in agenda

WP6: presentazioni a scuole calcolo parallelo

Veloce discussione su possibilità di partecipare alla scuola di calcolo parallelo di bertinoro presentando il progetto, alcuni risultati sul porting di applicazioni su schede low power e se possibile organizzare alcuni esercizi sulle schede stesse, magari portate in sede corso. Da approfondire con Mauro Morandin.

Partecipazione CCR Frascati

Sessioni di interesse mercoledì 27 pomeriggio e giovedi' 28 al mattino

Per il 27 l'agenda è pronta, resta da decidere alcuni speaker e scrivere un paio di abstract.

Considerazioni finali

Per la prima parte del progetto PM1-6 siamo sostanzialmente in linea con quanto scritto nel proposal – in particolare per quanto riguarda la creazione di benchmark e porting di applicazioni su piattaforme low power abbiamo fatto molto più di quanto inizialmente previsto.

Per il periodo PM6-12 resta l'incognita della disponibilità di development board basate su SoC con cpu a 64bit che è ormai un requirement stringente in particolare per le applicazioni LHC.

Resta quindi ancora incerto se riusciremo a costruire il cluster cnaf a 64bit prima della fine dell'anno così come scritto nel proposal. Potremmo costruirlo a 32bit ma si è deciso di attendere l'evoluzione del mercato a 64bit.

Data però la disponibilità di hw interessante (switch pci-e con un numero elevato di linee) per il wp di interconnessioni a bassa latenza potremmo concentrarci su quello nella seconda parte dell'anno.

Prossimo meeting

Organizziamo per la settimana del 24 giugno una phone conference (doodle da preparare) con all'ordine del giorno:

- Preventivi 2015 (consideriamo come non approvato il progetto picolo)
- Eventuali richieste per spese non previste
- Eventuali richieste per partecipazioni a conferenze visto che stiamo sottemettendo in diversi a vari eventi i risultati ottenuti

Proviamo ad organizzare anche una phone conference con il rappresentante italiano di APRE quanto prima.